

Solid State TECHNOLOGY

Insights for Electronics Manufacturing

**Emerging Memories for
the Zettabyte Era**

P. 14

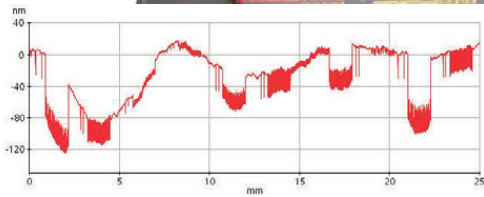
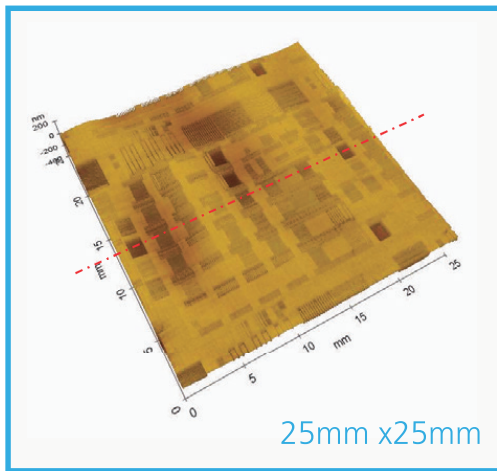
**AI Chips: Challenges
and Opportunities**

P. 18

**New Thinking Required
for Machine Learning**

P. 23

**Perfecting Yield with
Proactive Optimization** P.10

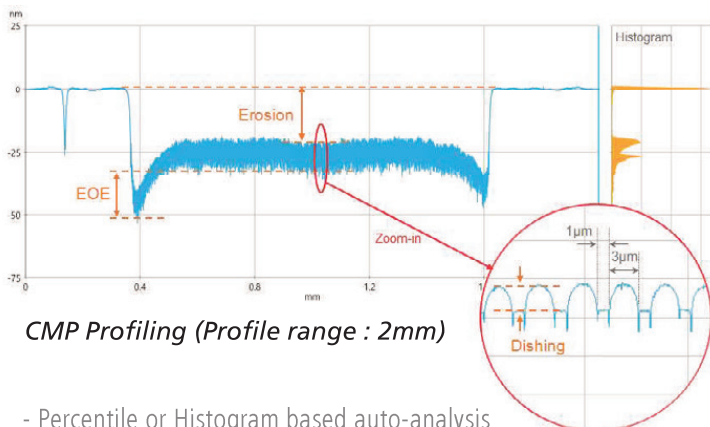


Long range profiling for CMP characterization

Park NX-Wafer

Low Noise, High Throughput Atomic Force Microscope with Automatic Defect Review & Atomic Force Profiler

- Low noise atomic force profiler for more accurate CMP profile measurements
- Sub-angstrom surface roughness measurements
- Fully automated AFM solution for defect imaging and analysis
- Capable of scanning 300 mm wafers
- Can improve defect review productivity by up to 1000%



- Percentile or Histogram based auto-analysis
- Erosion, EOE(Edge-Over-Erosion) and Dishing



www.parksystems.com
inquiry@parksystems.com

SEMICON EUROPA

November 13-16, 2018 - Munich / Germany

Park
SYSTEMS

Park Booth
#A4-644

Caption: Proactive yield perfection (PYP) is a comprehensive systems-level approach to perfecting yield through detailed surveillance and sophisticated modeling. Source: Rudolph Technologies

FEATURES

10

YIELD | Perfecting yield with proactive optimization throughout the process and across the supply chain

Increasingly dispersed and complex supply chains require a proactive, integrated, systems-level approach to optimizing yields.

Prasad Bachiraju, Rudolph Technologies, Inc., Wilmington, Mass.

14

MEMORY | Emerging memories for the zettabyte era

The scaling of traditional memories such as SRAM, DRAM and Flash is no longer following the data growth rate, especially in terms of energy and speed.

Gouri Sankar Kar and Arnaud Furnemont, imec, Leuven, Belgium

18

ARTIFICIAL INTELLIGENCE | AI chips: Challenges and opportunities

To get to the next level in performance/Watt, innovations being researched at the AI chip level include low precision, analog and resistive computing.

Pete Singer, Editor-in-Chief

23

ARTIFICIAL INTELLIGENCE | New thinking required for machine learning

Speakers argue the semiconductor community thus far has not been doing enough to enable machine intelligence.

Dave Lammers, Contributing Editor

25

MATERIALS | There's still plenty of room at the bottom: Isotopically pure materials for when every atom counts

When different isotopes of atoms have significantly different properties, the ability to create isotopically pure materials becomes essential.

Dr. Paul Stockman, Linde Electronics, Taipei, Taiwan

27

YIELD OPTIMIZATION | Dynamic Fault Detection: Utilizing AI and IoT to revolutionize manufacturing

A new approach in Fault Detection and Classification (FDC) allows engineers to uncover issues more thoroughly and accurately by taking advantage of full sensor traces.

Tom Ho and Stewart Chalmers, BisTEL, Santa Clara, CA

COLUMNS

- 2 **Editorial** | A new era of growth
Pete Singer, Editor-in-Chief
- 8 **Packaging** | Samsung at ECTC: Emphasis on warpage control
Phil Garrou, Contributing Editor
- 31 **Industry Forum** | Sensors in the new age of the car
Richard Dixon, Senior Principal Analyst, Sensors, IHS Markit

DEPARTMENTS

- 4 Web Exclusives
- 5 News
- 30 Ad Index



A new era of growth

The semiconductor industry is in a new era of growth, driven by a diverse array of applications and new computing architectures.

Just one indicator of the growth: Total wafer shipments in 2018 year are expected to eclipse the all-time market high set in 2017 and continue to reach record levels through 2021, according to SEMI's recent semiconductor industry annual silicon shipment forecast.

Much of this growth will come from the need for better connectivity and more intelligent data analysis using artificial intelligence (AI). AI represents a market opportunity \$2 trillion of on top of the existing \$1.5-2B information technology industry. According to IBM's John E. Kelly, "This is the era that's going to power our semiconductor industry forward. The number of opportunities is enormous." See AI chips: Challenges and opportunities on pg. 23.

Making AI semiconductor engines will require a wildly innovative range of new materials, equipment, and design methodologies. Moore's Law seems to be slowing, so chips designed for AI and more traditional applications will likely include advanced packaging/heterogeneous integration (think chip stacking), and silicon photonics.

In addition to AI chips from traditional IC companies such as Intel, IBM and Qualcomm, more than 45 start-ups are working to develop new AI chips, with VC investments of more than \$1.5B -- at least five of them have raised more than \$100 million from investors. Tech giants such as Google, Facebook, Microsoft, Amazon, Baidu and Alibaba are also developing AI chips.

AI will be paired with new computational methods (think non Von Neumann), such as neuromorphic methods that mimic how the brain works, and quantum computing.

Faster communication with higher bandwidth will be required. 5G wireless communication is coming, as is improved WiFi, near-field communication, Bluetooth and satellite communication.

Huge opportunities exist in automotive electronics, as autonomous driving moves closer to reality. Again, AI will play a critical role. Virtual reality will be combined with AI to create a truly immersive experience that mankind has never experienced.

Semiconductors and AI will play an increasingly important role in the healthcare industry, as diagnostic tools and patient monitoring.

A good example of how this all will impact the semiconductor industry moving forward can be found in DARPA's Electronics Resurgence Initiative (ERI). ERI calls for innovative new approaches to microsystems materials, designs, and architectures. The President's budget for FY19 includes continued annual investments of \$300 million over the next five years for ERI's research efforts--potentially upwards of \$1.5 billion over the initiative's lifetime.

In one project in the initiative, Applied, Arm and Symetrix will work together to develop a coprorrelated electron switch. "Most of the materials we are studying in the context of this research should, according to classical theories, conduct electrons, but thanks to correlation they can also exhibit insulating properties under certain conditions due to electron-electron interactions that don't happen according to traditional band theory," notes David Thompson of Applied Materials. "Welcome to the weird and wonderful world of quantum matter!" he adds.

—Pete Singer, Editor-in-Chief

Solid State TECHNOLOGY

Pete Singer, Editor-in-Chief
Ph: 978.470.1806,
psinger@extensionmedia.com

Shannon Davis, Editor, Digital Media
Ph: 603.547.5309
sdavis@extensionmedia.com

Ed Korczynski, Senior Technical Editor,
edk@extensionmedia.com

Dave Lammers, Contributing Editor

Phil Garrou, Contributing Editor

Dick James, Contributing Editor

Vivek Bakshi, Contributing Editor

CREATIVE/PRODUCTION/ONLINE

Marjorie Sharp, Production Traffic
Coordinator

Nicky Jacobson, Senior Graphic Designer

Slava Dotsenko, Senior Web Developer

MARKETING/CIRCULATION

Jenna Johnson,
jjohnson@extensionmedia.com

CORPORATE OFFICERS

Extension Media, LLC

Vince Ridley, President and Publisher
vridley@extensionmedia.com

Clair Bright, Vice President and Publisher
Embedded Electronics Media Group
cbright@extensionmedia.com

For subscription inquiries:

Tel: 847.559.7500; Fax: 847.291.4816;
Customer Service e-mail: sst@omeda.com;
Subscribe: www.sst-subscribe.com

Solid State Technology is published eight times a year by Extension Media LLC, 1786 Street, San Francisco, CA 94107. Copyright © 2018 by Extension Media LLC. All rights reserved. Printed in the U.S.

OCTOBER 2018 VOL. 61 NO. 7 • **Solid State Technology** ©2018 (ISSN 0038-111X) **Subscriptions:** Domestic: one year: \$258.00, two years: \$413.00; one year Canada/Mexico: \$360.00, two years: \$573.00; one-year international airmail: \$434.00, two years: \$691.00; Single copy price: \$15.00 in the US, and \$20.00 elsewhere. Digital distribution: \$130.00. You will continue to receive your subscription free of charge. This fee is only for air mail delivery. Address correspondence regarding subscriptions (including change of address) to: *Solid State Technology*, 1786 18th Street, San Francisco, CA 94107-2343. (8 am – 5 pm, PST).

**Extension
MEDIA**

1786 18th Street
San Francisco, CA 94107

SEMICONDUCTORS DRIVE SMART

13-16 Nov 2018, Messe München, Munich, Germany

This year SEMICON Europa will be the strongest single event for electronics manufacturing in Europe and is broadening the range of attendees across the electronics supply chain. The event covers the areas of Materials, Semiconductors, Frontend and Back-end Manufacturing, Advanced Packaging, MEMS/Sensors, Power and Flexible Electronics, and Automotive.

We connect the breadth of the entire electronics supply chain by including applications such as the Internet of Things, Artificial Intelligence, Machine Learning, and other adjacent markets.

CONFERENCES:

Automotive Forum
22nd Fab Management Forum
2018 FLEX Europe - Be Flexible
Advanced Packaging Conference
Strategic Materials Conference

EXHIBITION AND FREE PROGRAMS:

Showfloor (Hall A4)
TechARENA
TechLOUNGE
Talent & Leadership in the Digital Economy

EXECUTIVE NETWORKING EVENTS:

Executive Keynote Opening
SEMI Networking Night

KEY FACTS

More than ...



35% visitors represent engineering job functions



70% involved in purchasing decisions



33% are executives and senior managers



8,700 visitors from 85 countries registered in 2017

INCREASE YOUR EXPOSURE!

Contact us to **become a sponsor!**

SEMI Europe Sales:
Denada Hodaj, dhodaj@semi.org

To **reserve your booth** contact
SEMI Europe Tradeshow Operations:
SEMICONEuropa@semi.org

REGISTER NOW!



Web Exclusives

Automating 200mm semiconductor fabs to meet growing demand

SEMI met with Heinz Martin Esser, managing director at Fabmatics GmbH, to discuss how existing 200mm semiconductor fabs can master the challenges of a 24x7 production under highest cost and quality pressure by implementing intralogistics automation solutions. The two spoke ahead to his presentation at the Fab Management Forum at SEMICON Europa 2018, 13-16, November 2018, in Munich, Germany.

<https://bit.ly/2yTHO2w>

Insights from the Leading Edge: IC history for the younger generation

The IC industry started out like a poker championship tournament. Hundreds of players, through the years, put up their entry fee to compete (i.e. paying for their fabs) and the competition began.

<https://bit.ly/2PjeYj7>

How will graphene, the 2D wonder material, change the semiconductor industry?

Materials innovation has always been vital to the semiconductor industry. In the past, it was high- κ gate dielectrics. Today, Cobalt is seen as a replacement for Tungsten in middle-of-line (MOL) contacts. What materials innovation will the future bring? A likely answer is graphene, the wonder material discovered in 2004.

<https://bit.ly/2Ply20b>

Power subsystems: Surging on a wave of vacuum processing equipment

Process power and reactive gas subsystems for semiconductor manufacturing equipment have grown at a CAGR of 21% since 2013. The segment growth is considerably above the critical subsystems industry average of 9.5% and is attributable to higher demand for vacuum processing equipment over the period.

<https://bit.ly/2E9u49U>



IBM's Jeff Welser to Keynote The ConFab 2019

AI was a big focus on The ConFab and 2018 and we will continue that theme in 2019 with a keynote talk by IBM's Jeff Welser. The ConFab 2019 will return to The Cosmopolitan of Las Vegas on May 14-17. In 2018, AI and other leading technologies were discussed by speakers from IBM, Google, Nvidia, HERE Technologies, Silicon Catalyst, TechInsights, Siemens and Qorvo, among many others.

<https://bit.ly/2Ilj6Xj>

Ruthenium nanolayers are ferromagnetic at RT

Researchers from Intel Corporation and the University of Minnesota and the University of Wisconsin have shown that strained atom-scale films of pure ruthenium (Ru) metal exhibit ferromagnetism at room temperature, opening up the possibility of using the material to build novel magnetic random access memory (MRAM) devices. (From *SemiMD.com*)

<https://bit.ly/2MsM1UH>

SEMI calls for exclusion process for products in new China tariff list

SEMI joined a coalition of business groups in calling for Ambassador Robert Lighthizer, U.S. Trade Representative, to enact an exclusion process for the most recent tranche of tariffs on \$200 billion in goods imported from China.

<https://bit.ly/2OdCCKD>

Toshiba Memory and Western Digital celebrate the opening of Fab 6 and Memory R&D Center at Yokkaichi, Japan

Toshiba Memory Corporation and Western Digital Corporation (NASDAQ:WDC) yesterday celebrated the opening of a new semiconductor fabrication facility, Fab 6, and the Memory R&D Center, at Yokkaichi operations in Mie Prefecture, Japan.

<https://bit.ly/2yl5xsd>

worldnews

EUROPE - **Leti** and **EFI** launched a project to improve reliability and speed of low-cost electronic devices for autos.

ASIA - **MagnaChip** will host Foundry Technology Symposium in Shenzhen, China in November 2018.

USA - **GLOBALFOUNDRIES** extended FinFET offering with new features to enable tomorrow's intelligent systems.

EUROPE - **STMicroelectronics** and **Leti** announced their cooperation to industrialize GaN (Gallium Nitride)-on-Silicon technologies for power switching devices.

USA - **Lam Research** named its 2018 Supplier Excellence Award recipients.

ASIA - **Air Products** announced it has been awarded by **Samsung Electronics** additional gaseous nitrogen and hydrogen supply to its semiconductor fab in Giheung, South Korea.

USA - **Imagination** and **GLOBALFOUNDRIES** announced a collaboration to deliver ultra-low-power connectivity solutions for IoT applications.

ASIA - **Amkor** opened a new semiconductor package manufacturing and test plant in Taiwan.

USA - **SUNY Poly** professor was awarded \$500,000 National Science Foundation grant to develop next-generation memory technology.

USA - **Keysight Technologies'** 3D planar electromagnetic simulator certified for **GLOBALFOUNDRIES** 22FDX process technology.

New fabs invest over \$220B; 2019 to mark all-time spending high

Global fab equipment spending will increase 14 percent this year to US\$62.8 billion and is expected to rise 7.5 percent, to US\$67.5 billion, in 2019, marking the fourth consecutive year of spending growth and the highest investment year for fab equipment in the history of the industry, according to the latest World Fab Forecast Report published by SEMI. Investments in new fab construction are also nearing a record with a fourth consecutive year of growth predicted and capital outlays next year approaching US\$17 billion.

Investments for fab technology and product upgrades, as well as for additional capacity, will grow as the emergence of numerous new fabs significantly increases equipment demand, the forecast shows. The World Fab Forecast Report currently tracks 78 new fabs and lines that have or will start construction between 2017 to 2020 (with various probabilities) and will eventually require more US\$220 billion in fab equipment (Figure 1). Construction spending for these fabs and lines is expected to reach US\$53 billion during this period.

Korea is projected to lead other regions in fab equipment investments with US\$63 billion, US\$1 billion more than second-place China. Taiwan is expected to claim the third spot at US\$40 billion, followed by Japan at US\$22 billion and the Americas at US\$15 billion. Europe and Southeast Asia will share sixth

place, with investments totaling US\$8 billion each. Fully 60 percent of these fabs will serve the Memory sector (the lion's share will be 3D NAND), and a third will go to Foundry.

Of the 78 fab construction projects starting construction between 2017 and 2020, 59 began construction in the first two years (2017 and 2018), while 19 are expected to begin in the last two years (2019 and 2020) of the tracking period.

Equipping a new fab typically takes one to one and a half years, though some fabs take two years and others longer, depending on various factors as such the company, fab size, product type and region. Approximately half of the projected US\$220 billion will be spent from 2017 and 2020, with less than 10 percent invested in 2017 and 2018, nearly 40 percent in 2019 and 2020, and the rest after 2020.

While the US\$220 billion estimate is based on current insights of known and announced fab plans, total spending could exceed this level as many companies continue to announce plans for new fabs. Since the last quarterly publication of the report published last quarter, 18 new records – all new fabs – have been added to the forecast. Up-to-date and detailed analysis, with a bottoms-up approach, is available by subscribing to SEMI's World Fab Forecast Report.

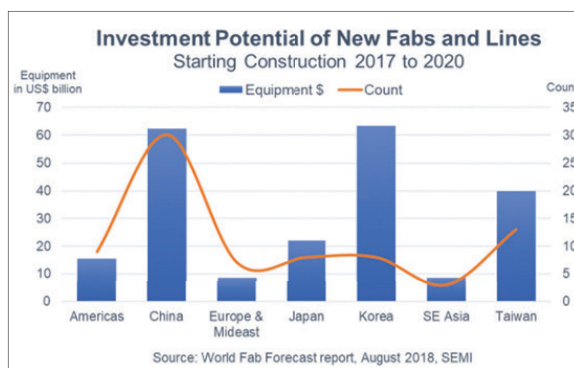
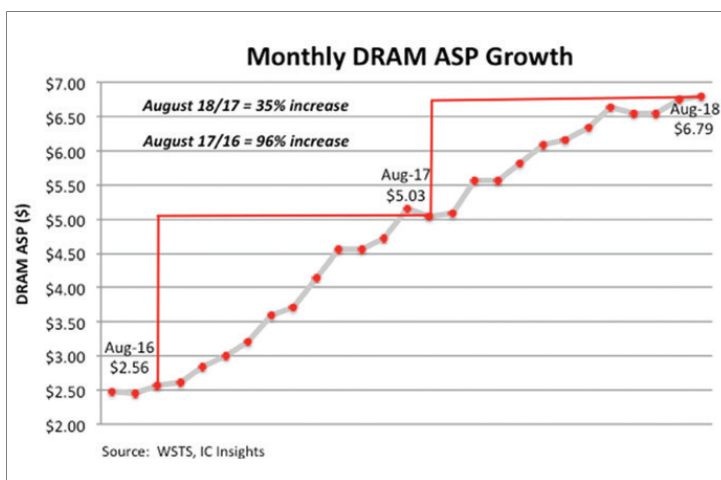


FIGURE 1. Shows the investment potential of new fabs and lines starting construction between 2017 and 2020.

Since its June 1 publication, more than 340 updates have been made to the World Fab Forecast. The report now includes more than 1,200 records of current and future front-end semiconductor facilities from high-volume production to research and development. The report covers data and predictions through 2019, including milestones, detailed investments by quarter, product types, technology nodes and capacities down to fab and project level. ◀▶

DRAM market braces for slower growth

In its September Update to The 2018 McClean Report, IC Insights discloses that over the past two years, DRAM manufacturers have been operating their memory fabs at nearly full capacity, which has resulted in steadily increasing DRAM prices and sizable profits for suppliers along the way. Figure 1 shows that the DRAM average selling price (ASP) reached \$6.79 in August 2018, a 165% increase from two years earlier in August of 2016. Although the DRAM ASP growth rate has slowed this year compared to last, it has remained on a solid upward trajectory through the first eight months of 2018.



The DRAM market is known for being very cyclical and after experiencing strong gains for two years, historical precedence now strongly suggests that the DRAM ASP (and market) will soon begin trending downward. One indicator suggesting that the DRAM ASP is on the verge of decline is back-to-back years of huge increases in DRAM capital spending to expand or add new fab capacity (Figure 2). DRAM capital spending jumped 81% to \$16.3 billion in 2017 and is expected to climb another 40% to \$22.9 billion this year. Capex spending at these levels would normally lead to an overwhelming flood of new capacity and a subsequent rapid decline in prices.

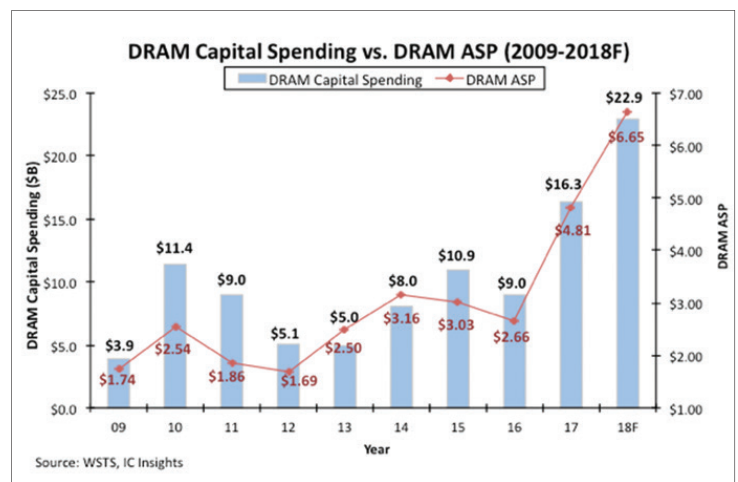
However, what is slightly different this time around is that big productivity gains normally associated with significant spending upgrades are much less at the sub-20nm process node now being used by the top DRAM suppliers as compared to the gains seen in previous generations.

At its Analyst Day event held earlier this year, Micron presented figures showing that manufacturing DRAM at the sub-20nm node required a 35% increase in the number of mask levels, a 110% increase in the number of non-lithography steps per critical mask level, and 80% more cleanroom

space per wafer out since more equipment—each piece with a larger footprint than its previous generation—is required to fabricate $\leq 20\text{nm}$ devices. Bit volume increases that previously averaged around 50% following the transition to a smaller technology node, are a fraction of that amount at the $\leq 20\text{nm}$ node. The net result is suppliers must invest much more money for a smaller increase in bit volume output. So, the recent uptick in capital spending, while extraordinary, may not result in a similar amount of excess capacity, as has been the case in the past.

As seen in Figure 2, the DRAM ASP is forecast to rise 38% in 2018 to \$6.65, but IC Insights forecasts that DRAM market growth will cool as additional capacity is brought online and supply constraints begin to ease. (It is worth mentioning that Samsung and SK Hynix in 3Q18 reportedly deferred some of their expansion plans in light of expected softening in customer demand.)

Of course, a wildcard in the DRAM market is the role and impact that the startup Chinese companies will have over the next few years. It is estimated that China accounts for approximately 40% of the DRAM market and approximately 35% of the flash memory market.



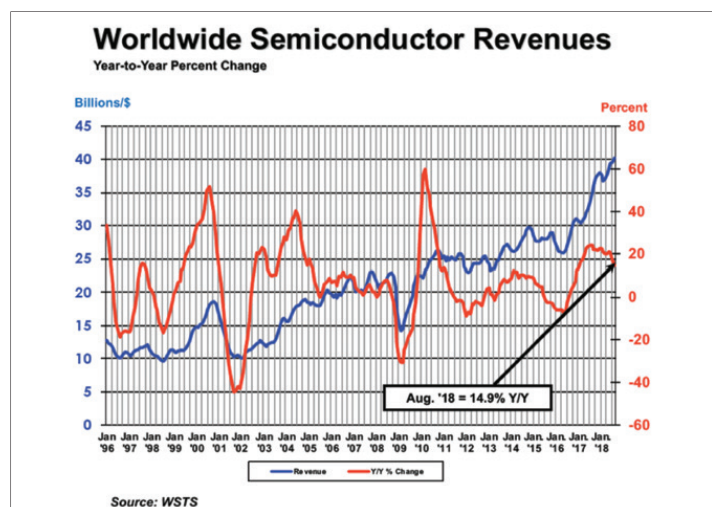
At least two Chinese IC suppliers, Innotron and JHICC, are set to participate in this year's DRAM market. Although China's capacity and manufacturing processes will not initially rival those from Samsung, SK Hynix, or Micron, it will be interesting to see how well the country's startup companies perform and whether they will exist to serve China's national interests only or if they will expand to serve global needs. ◀

Global semiconductor sales increase 14.9% year-to-year in August

The Semiconductor Industry Association (SIA), representing U.S. leadership in semiconductor manufacturing, design, and research, announced worldwide sales of semiconductors reached \$40.16 billion for the month of August 2018, an increase of 14.9 percent compared to the August 2017 total of \$34.96 billion. Global sales in August 2018 were 1.7 percent higher than the July 2018 total of \$39.49 billion. All monthly sales numbers are compiled by the World Semiconductor Trade Statistics (WSTS) organization and represent a three-month moving average.

"Global semiconductor sales continued to bound upward in August, easily outperforming sales from last August and narrowly surpassing last month's total," said John Neuffer, president and CEO, Semiconductor Industry Association. "While year-to-year growth has moderated somewhat in recent months, sales remain strong across every major semiconductor product category and regional market, with the China and Americas markets standing out with the largest year-year growth."

Regionally, sales increased compared to August 2017 in China (27.3 percent), the Americas (15.0 percent), Europe (9.5 percent), Japan (8.4 percent), and Asia Pacific/All Other (4.7 percent). Sales were up compared to last month in China (2.1



percent), the Americas (3.6 percent), and Asia Pacific/All Other (1.3 percent), and decreased slightly in Japan (-0.1 percent), and Europe (-1.4 percent).

For comprehensive monthly semiconductor sales data and detailed WSTS Forecasts, consider purchasing the WSTS Subscription Package. For detailed data on the global and U.S. semiconductor industry and market, consider purchasing the 2018 SIA Databook. ◀



HETEROGENEOUS INTEGRATION: THE PATH FORWARD

REALIZING THE COST AND PERFORMANCE BENEFITS

12.5.2018

WEDNESDAY, DECEMBER 5, 2018 | MILPITAS, CALIFORNIA

SPONSORS AND EXHIBITORS

SMART
MICROSYSTEMS™



Promex
Microelectronics Assembly Technologies

Mentor®
A Siemens Business

REGISTER ONLINE TODAY AT WWW.MEPTEC.ORG

Samsung at ECTC: Emphasis on warpage control






PHIL GARROU,
Contributing Editor

At the 2018 ECTC, Samsung presented several papers on their advanced packaging activities. In one paper, Samsung teamed with SUNY Binghamton to discuss “design Guidelines of 2.5D Package with Emphasis on Warpage Control and Thermal Management.” A 2.5D Package is composed of many material sets and in general its size is larger than conventional single chip packages. The CTE of substrate is a well-known factor to control warpage in a single die packages. However, the existence of another layer (interposer) makes the problem more complicated. Optimization of the material sets, which include lid, EMC, chip, Interposer and geometric factors, are essential. From their modeling studies it is clear that substrate CTE is more influential than other criteria.

Area ratio of lid attach, lid thickness, EMC CTE and substrate CTE are major factors influencing for warpage. For thermal management, EMC coverage on top of the chip, (cooling) fan speed, and conductivity of TIM are the major factors that affect thermal resistance.

In their presentation on “Low Cost Si-less RDL Interposer Pkg for High Performance Computing Applications” In this presentation a concept for a Si-less redistribution layer is described for server/HPC applications and warpage behavior, electrical performance and reliability of the RDL interposer package were evaluated.

Table 1. Comparison of 2.5D Si interposer and Si-less RDL interposer

Test	2.5D Si interposer	Si-less RDL interposer	
		Wafer level	Panel level
L/S	< 1 μm	< 2~5 μm	2 μm target
Cost	High	Middle	Low
Productivity	Low (1x) 	Low (1x) 	High (~3.5x) 
Application	High-end packages with fine L/S	Low cost and large size packages with wider L/S	

Si-interposer have attracted attention for high end server products due to high electrical performance at low power consumption. The key barrier of Si-interposer adoption, utilizing TSV, is high manufacturing cost for large interposer sizes. They suggest a Si-less redistribution layer (RDL) interposer platform for high performance applications as a low cost package solution.

The Table compares 2.5D Si interposer technology to wafer level and panel level RDL interposers.

The fabrication process flow of RDL interposer package is classified into six main steps as summarized in the fig below. RDL formation, multi-chip bonding on RDL, encapsulation, chip exposure, solder ball attachment, and interposer assembly on PCB. The most challenging aspect of the assembly is reportedly the warpage control of interposer packages, due to the large size and multichips.

Samsung claims that the RDL interposer package has the advantage of lower manufacturing cost over Si-interposer by replacing TSV with RDL. Their results showed that RDL interposer warpage is more controllable than Si-interposer at room and high temperature by the optimization of design, process condition and material selection. Their test structure, a RDL interposer package whose size is larger than 3000mm² included four HBMs and one ASIC chip, was successfully fabricated and they determined that the electrical loss of RDL interposer was lower than Si interposer case. Mechanical simulation showed RDL interposer reduced joint stress by 34% compared to Si interposer. They predict that RDL interposer tech will become one of the most promising solutions for low cost and large size packages in the near future if “...fine patterning technology is developed below L/S 2/2 μm .” ◀

Packaging





The ConFab 2019
Distinguished Opening Keynote

Dr. Jeffrey Welser,
Vice President, IBM Research – Almaden

May 14-17, 2019 at The Cosmopolitan of Las Vegas
theconfab.com

Perfecting yield with proactive optimization throughout the process and across the supply chain

PRASAD BACHIRAJU, Rudolph Technologies, Inc., Wilmington, Mass.

Increasingly dispersed and complex supply chains require a proactive, integrated, systems-level approach to optimizing yields.

After decades of R&D, two emerging memory types – the phase change memory-based 3D Xpoint, co-developed by Intel and Micron, and the embedded spin-torque transfer magnetic RAM (e-MRAM) from several foundries – are now coming to the market. One point of interest is that neither memory type relies on the charge-based SRAM and DRAM memory technologies that increasingly face difficult scaling challenges. Another is that both have inherent performance advantages that could extend their uses for decades to come.

As IC manufacturers adopt advanced packaging processes and heterogeneous, multi-chip integration schemes to feed ever-greater consumer demand for more computing power in smaller and smaller spaces, their supply chains have become increasingly complex. A fab-wide view of the process, not so long ago the holy grail of yield management systems, now seems quaintly inadequate. Critical processes that affect finished product yields now occur in different facilities running diverse processes at locations spread around the world. A packaged electronics module may combine microprocessors, memory, MEMS sensors and RF communications, all from different fabs and each with its own particular history. Optimizing yields from such a complicated supply chain requires access to individual component genealogy that includes detailed knowledge of process events, equipment malfunctions and operational parameters, and much more. The latest generation of yield optimization systems consolidates this data – tool deep and supply-chain wide – in a monolithic database, providing end-to-end, die-level traceability, from bare wafer to final module test. Specialized algorithms, designed for “big data” but based on intimate knowledge of semiconductor manufacturing processes, can find hidden correlations among param-

eters, events and conditions that guide engineers to the root causes of yield losses and ultimately deliver increased fab productivity, higher process yields, and more reliable products.

Proactive Yield Perfection

Proactive yield perfection (PYP) is a comprehensive systems-level approach to perfecting yield through detailed surveillance and sophisticated modeling that identifies actual or potential root causes of excursions and establishes monitoring mechanisms to anticipate and proactively address problems before they result in yield loss. PYP is a logical extension of long-standing yield management practices that comprehends the dispersal and growing complexity of the manufacturing process and combines information from conventional defect detection, yield analysis, automated process control, and fault detection and classification techniques. It addresses

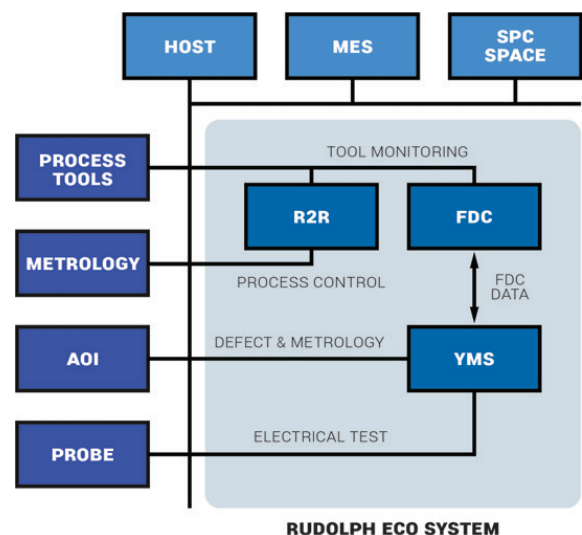


FIGURE 1. The PYP ecosystem

two major obstacles identified within the industry: providing access to data across the fab and supply chain and integrating that data into device-level genealogy. PYP's ability to correct small problems early is increasingly valuable in a complex supply chain where each step represents a considerable additional investment and a flawed finished module results in the costly loss of multiple component devices.

PYP collects data in a single database that integrates vital product parameter data from every die at each step in the process with performance and condition data from all tools, factories and providers in the supply chain. It provides unprecedented visibility of the entire manufacturing process from design, through wafer fab, test, assembly and packaging. Consolidating the data in a single, internally consistent database eliminates the considerable time often consumed in simply locating and aligning data stored in disparate databases at multiple facilities, and allows analytical routines to find correlations among widely separated observations that are otherwise invisible (**FIGURE 1**).

Genealogy incorporates traditional device-level traceability of every die, but also provides access to all information available from sensors on the tool (temperature, pressure), process events (e.g. lot-to-lot changes and queue times), equipment events (e.g. alarms and preventive maintenance), changes in process configurations (e.g. specifications and recipes), and any other event or condition captured in the database for sophisticated analysis. Ready access

to device genealogy allows analysts to trace back from failed devices to find commonalities that identify root causes, and trace forward from causes and events to find other device at risk of failure.

3x longer TC life.
30% lower annual spend.



EtchDefender™
TECHNOLOGY

Only from Conax Technologies

EtchDefender™ technology extends the life of quartz thermocouple sheaths in ASM® EPSILON® epitaxial reactors up to **3X longer and reduces your annual spend by 30%.**



Ideas. Solutions. Success.

+1 800 223 2389 | conaxtechnologies.com/ss

ASM® and EPSILON® are registered trademarks of ASM® International. Neither Conax Technologies nor its products are affiliated with, approved by or sponsored by ASM® International.



FIGURE 2. Numerous locations worldwide feed information into a single big-data cluster where multiple software solutions operate on the pre-aligned and internally consistent data set.

Mobile communications

A leading global supplier of mobile communications products has implemented the full PYP suite across their supply chain (**FIGURE 2**), which includes three separately located front-end fabs, a fabless design facility, outsourced manufacturing (foundries), and outsourced assembly and test (OSAT). Finished modules integrate data processing, data storage, RF communications, power management, analog sensing, and other functions using multiple die and components fabricated at various facilities around the world. The PYP suite comprises fault detection and classification, yield management, defect detection and classification, and run-to-run automated process control, including configurable dashboard displays that provide

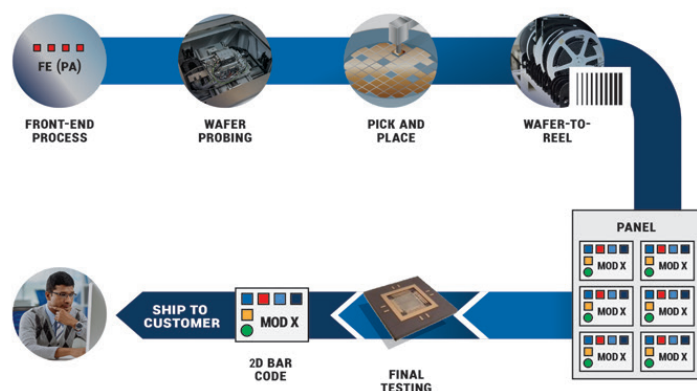


FIGURE 3. In a typical process flow, die move from wafer (front-end and back-end wafer fab), to reel, to panel, to final test

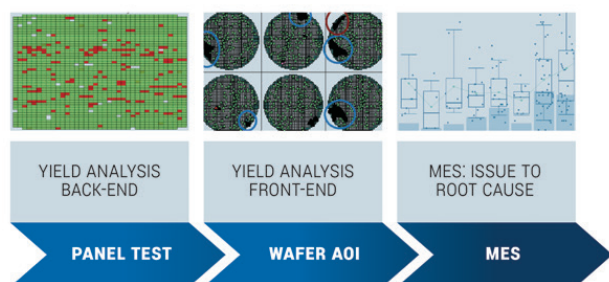


FIGURE 4. A potential loss of hundreds of thousands of dollars due to final test on panel (left) was traced back to over a thousand wafers. Through yield analysis, an SPR pattern signaled wafers with edge defect clusters (center). Finally, the root cause is understood through MES analysis (right).

interactive drill-down reports and scheduled user-definable reports. The process flow includes front-end wafer-based processing, with singulated die then transferred to tape and subsequently to rectangular panels for back-end processing (**FIGURE 3**).

In one instance (**FIGURE 4**), panel mounted die were failing in back-end processing, causing yield losses valued at hundreds of thousands of dollars. The failed die were traced back to over a thousand wafers, and analysis of those wafers, using spatial pattern recognition, revealed defect clusters near the wafer edges. Further analysis of integrated MES (wafer process history) data traced the clusters to a defective tool that leaked etching solution onto wafers. Engineers are currently evaluating sensor data from the tool in an effort to identify a signal that will permit proactive intervention to prevent similar losses in the future.

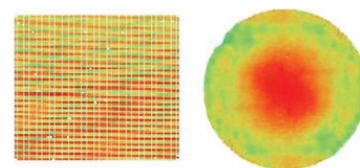


FIGURE 5. The bad die (red strips) on the panel (left) were traced back the centers of the original wafers.

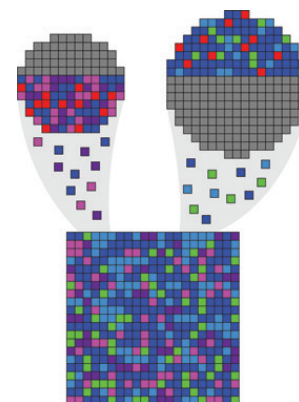


FIGURE 5. A single panel contains different filter components from multiple (differently-sized) wafers. Die-level traceability allows engineers to relate final test results to inspection and metrology data collected during wafer processing. Die can be traced back to their individual location on the original wafer, permitting associations with location-specific data such as defect patterns discovered by SPR.

In another case, failed die on panels in the back-end showed a characteristic “strip” signature (**FIGURE 5**). Tracing the die backward revealed a front-end a process issue with the failing die all originating from locations near the center of the wafer.

A final example from this manufacturer occurred in the back-end where die containing various filter technologies fabricated on a wafers of different sizes are combined on a panel substrate (**FIGURE 6**). In this case, the customer gained additional insights regarding the origin of the die being assembled and was able to evaluate how shifts in performance parameters impacted the final module product. This resulted in better matching of parts in the pre-assembly process and a tighter distribution of performance parameters in the outgoing modules.

Automotive electronics

A leading global supplier of electronic systems to the automotive market manufactures finished modules containing multiple ICs from various suppliers and facilities. Highest reliability with component failure rates at parts-per-billion levels is an absolute requirement because the health and safety of millions of drivers may be jeopardized by a defective product. Limiting these risks, and the associated financial liability, through fast root-cause analysis of in-house test failures and field returns and rapid identification of process drift or step function changes are critical needs. Tuning the process to improve yields while preserving critical reliability is an equally important economic concern.

The final product is a multi-chip module containing microelectromechanical system (MEMS)-based sensors from one supplier and application-specific integrated circuits (ASIC) from another. These component parts are functionally tested prior to being sent to an assembly facility where they are attached to a common carrier and packaged in a sealed module, which is then retested to verify functionality of the completed assembly. PYP software collects data across the entire supply chain. Devices from a single wafer lot are ultimately split and mixed among many modules. Using a commonality analysis, engineers can quickly identify die that share a similar risk and track them to their ultimate dispositions in finished modules. The tracking is not limited to wafer level. For instance, it might be used to find only those die located on the straight-line extension of a known crack or within at-risk regions identified by spatial pattern recognition (SPR). Such information is critical in issuing a recall of at-risk parts.

In this case, die that were known-good at wafer test were failing after assembly. Tracing back to wafer level the customer determined that all affected modules were assembled within a short time period of one another. Further investigation found that the packaging process was affecting peak-to-peak voltage at final test. The customer was not able to modify the assembly process, but they were able to eliminate the final test losses by tuning wafer probe specifications to eliminate die at risk for damage in the assembly operation. Die-level

traceability across the supply chain, which allowed engineers to quickly and easily compare data sets on the same dies from wafer probe and final test, was key to achieving this solution.

Conclusion

Increasingly dispersed and complex supply chains require a proactive, integrated, systems-level approach to optimizing yields. PYP's ability to integrate data – sensor-deep and supply chain-wide – in a monolithic database streamlines analysis and finds relationships that are otherwise invisible. Die-level genealogy allows engineers to trace die histories backward to find root causes of failures and forward to identify other die similarly at-risk. The value of PYP-based solutions is multiplied by the substantial investments made at each step of the process and the high cost and potential financial liability associated with failed, multi-chip modules.

PRASAD BACHIRAJU is Director, Customer Solutions, Rudolph Technologies, Inc., Wilmington, Mass. ◀



Delivering Real-world Solutions

Nikon combines a strong history of engineering expertise with superior scanner technology and innovative alignment solutions to deliver exceptional manufacturing performance and productivity—now and for the future.

Nikon. Delivering Real-world Solutions.

See Nikon at SEMICON Europa 2018 - Visit booth A4706 to learn about the latest lithography solutions.

www.nikonprecision.com

Emerging memories for the zettabyte era

GOURI SANKAR KAR and **ARNAUD FURNEMONT**, imec, Leuven, Belgium

The scaling of traditional memories such as SRAM, DRAM and Flash is no longer following the data growth rate, especially in terms of energy and speed.

Every day, even every second, we produce massive amounts of data. With an estimated annual data growth rate of 1.2 to 1.4x (source: IDC's Data Age 2025 study, March 2017), the amount of digital data produced in the world will soon exceed 100 zettabyte. To grasp the meaning of this number – we would need to fill a soccer field with terabyte solid state drives (SSDs) stacked 28 meter high if we would want to store all this data. The data are partly generated through well-known applications such as Amazon, YouTube, Facebook or Netflix. But emerging IoT applications will make a significant contribution as well. Examples are the autonomous car (accounting for 4,000GB of data per day), the smart building (>275GB per day, per building) and the smart city (>1000TB per day, per city). Huge amounts of bandwidth are required to transport all this data – from the application to an edge node, then to a base station, and then to a data center – a challenge that is tackled by 5G and optical fiber technologies. Throughout this data flow, stringent requirements will be imposed on memory and storage – in terms of density, bandwidth, cost and energy.

Clever data mining, and reduced energy consumption

At some point in the flow of data transport, the generated data will need to be analyzed and converted into knowledge and wisdom by means of machine learning techniques. The exact point at which this will happen, will significantly impact the requirements on memory and storage. For example, if machine learning can be applied just after data generation, it can help relax the requirements. If, on the other hand, data is turned into wisdom later in the process, more raw data will need to be stored throughout the whole process.

The zettabyte era will also challenge the power that is consumed by the growing amount of data centers, for

processing, transporting and storing all the data. If we don't optimize the energy consumption for these operations, data centers worldwide may use almost 8000 terawatt-hours by 2030 (source: <https://www.labs.hpe.com/next-next/energy>). That's about the amount of electricity consumed by Europe, Africa and part of Asia today. In recent years, several technologies have been introduced in the data centers to address power and performance issues for storage, including, for example, the wide deployment of solid state drives since 2014, and the introduction of the first emerging memory technologies in 2017. But to be prepared for the zettabyte era, we will have to introduce novel non-volatile memories with unprecedented density and speed, and improved power consumption.

The slowdown of today's memory roadmap

Let's have a closer look at today's memory landscape (**FIGURE 1**). Close to the central processing unit (CPU), fast, volatile embedded static random access memories (SRAMs) are the dominant memories. Also on chip are the higher cache memories, mostly made in SRAM or embedded dynamic random access memory (DRAM) technologies. Off-chip, further away from the CPU, you will mainly find DRAM chips for the working memory, non-volatile Flash NAND memory chips for storage, and tapes for long-term archival applications. In general, memories located further away from the CPU are cheaper, slower, denser and less volatile.

For half a century, Moore's Law has driven cost improvement of memory technologies, and this has translated into a continuous increase of the memory density. However, despite large improvements in memory density, only storage density (Flash NAND devices and tapes) has truly kept pace with the data growth rate. With the transition from NAND to 3D-NAND devices, density improvement for this storage class is however expected to slow down as well, and go below the data growth rate soon.

Unlike in logic development – which has always been driven by improvements in cost and device physics – improving the power/performance benefits for memory has barely been taken care of. As a result, energy reduction and speed improvement are far from following up the data growth rate, for both memory and storage devices (**FIGURE 2**).

Emerging technologies to the rescue?

To meet the memory requirements of the zettabyte era (i.e., improved density and speed, and reduced energy consumption), imec is exploring multiple emerging memory options, for standalone as well as for embedded applications (**FIGURE 3**). Options range from MRAM technologies for cache level applications, new ways for improving DRAM devices, emerging storage class memories to fill the gap between DRAM and NAND technologies, solutions for improving 3D-NAND storage devices, and a revolutionary solution for archival type of applications. Below, we review their status and challenges, and investigate whether these emerging memory roadmaps can cope with the zettabyte world.

MRAM technologies for embedded cache level applications

Spin transfer torque MRAM (STT-MRAM) technology has emerged as a candidate technology for replacing L3 cache embedded SRAM memories. It offers non-volatility, high density, high speed and low switching current. The core element of an STT-MRAM device is a magnetic tunnel junction in which a thin dielectric layer is sandwiched between a magnetic fixed layer and a magnetic free layer. Writing of the memory cell is performed by switching the magnetization for the free magnetic layer, by means of a current that is injected perpendicular into the magnetic tunnel junction.

Because of this geometry, the read and write operations are performed through the same path, and this challenges the reliability of the device. These reliability issues in combination with increased energy at sub-ns switching speeds make STT-MRAM memories unsuitable for replacing the faster L1/L2 cache SRAM memories.

An MRAM variant, the spin orbit torque MRAM (SOT-MRAM), can overcome these issues. In these devices, switching the free magnetic layer is done by injecting an in-plane current in an adjacent SOT layer, as such de-coupling the read and write path and improving the device endurance and stability. Imec recently demonstrated the

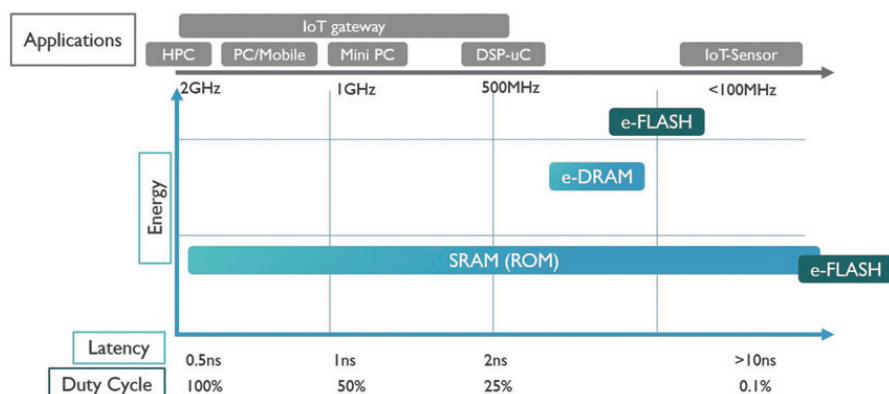


FIGURE 1. Application and performance space of the existing conventional memories (HPC = high performance computing; DSP = digital signal processing; ROM = read only memory; duty cycle = the proportion of time during which the memory device is operated).

	ENERGY REDUCTION FOR A GIVEN THROUGHPUT	DENSITY IMPROVEMENT	SPEED (AT DEVICE LEVEL)
CACHE (SRAM)	1.12x	1.15x	1.1x
MEMORY (DRAM)	1.1X Cs reduction, Vdd reduction	1.2X → 1.1X Slow down with C scaling	1x
STORAGE (FLASH)	1X	1.4X → 1.2X 3D log trend cannot last	1x performance increase from 2D to 3D
ARCHIVAL (TAPES)	1.1x	1.4X Doubling at each LTO node	1.1x

FIGURE 2. Growth rates for the main memory roadmaps; (red) only storage density has kept pace with the data growth rate.

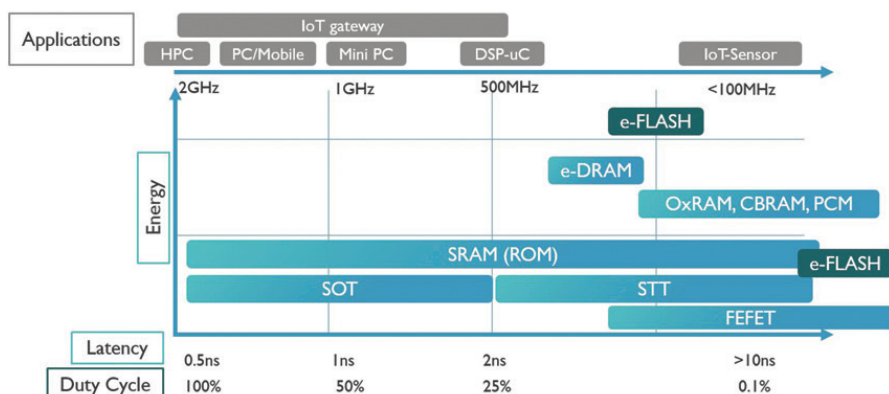


FIGURE 3. Memory research landscape at imec – application and performance space.

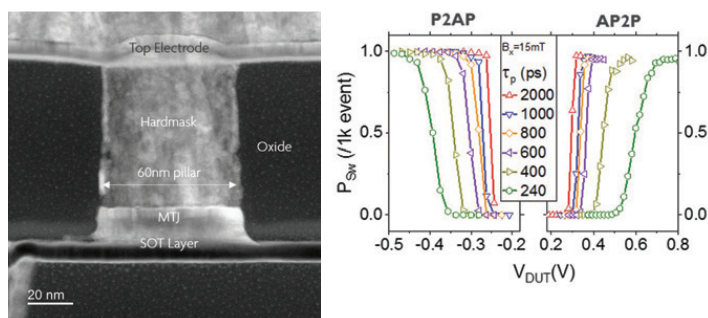


FIGURE 4. (Left) SOT-MRAM device, and (right) SOT switching distribution as a function of pulse voltage for various pulse lengths.

ability to fabricate state-of-the-art SOT-MRAM devices on 300mm wafers using CMOS compatible processes (**FIGURE 4**). The devices exhibited an unlimited endurance ($>5 \times 10^{10}$), fast switching speed (240ps), and power consumption as low as 300pJ. We also explore ways for further reducing energy consumption, by bringing down switching current and demonstrating field-free switching.

An imec view on DRAM scaling

DRAM is structurally a very simple type of memory. A DRAM memory cell consists of one transistor and one capacitor, that can be either charged or discharged. Traditionally, double-sided cylindrical capacitor structures have been used, with a dielectric material contained inside as well as outside the capacitor structure. However, when spacing becomes smaller, we need to increase the aspect ratio of these structures – since a certain amount of capacitor area is needed to maintain the memory's performance. For very large aspect ratios, these DRAM structures are fabricated at the limits of mechanical stability. The industry may therefore transition to a new capacitor architecture: the high-aspect ratio one-pillar architecture, with the dielectric film now only on the outside (**FIGURE 5**). This change may make it possible to use thicker films of higher-k dielectrics, which in turn will allow reducing the aspect ratio of the one-pillar structure and improving leakage. A new dielectric with these specifications is currently under development at imec.

Further down the road, we are investigating if we can place the peripheral logic directly under the array of capacitors and transistors. This logic circuitry controls how data is moved to and from the memory chip, and typically consumes considerable area. Today, the transistor of the DRAM memory cell is however built on silicon. To be able to move the peri logic underneath the DRAM array, we need to replace this transistor with a non-Si transistor that is back-end compatible. At imec, we are moving towards

a thin-film indium gallium zinc oxide (IGZO)-based transistor (**FIGURE 6**). This architecture is expected to give us one generation of Moore's law scaling. In addition, it will enable ultimate 3D DRAM integration as well.

Emerging memory and selector concepts for storage class memory

Storage class memory has been introduced to fill the gap between DRAM and NAND Flash memories in terms of latency, density, cost and performance. This new memory class should allow massive amounts of data to be accessed in a very short time. Most probably, more than one novel memory technology will be required to span the entire gap. The imec team explores several emerging technologies for storage class memory, including various cross-point-based architectures for the memory element, such as phase-change-RAM (PC-RAM), vacancy-modulated conductive oxide (VMCO), conductive bridging RAM (CB-RAM) and oxide RAM (OxRAM). When it comes to high-density applications, all these memory elements require two-terminal selector elements that connect serially with each of the memory elements. These selector elements suppress the sneak currents that run through the unselected cells in the cross-point array during memory operation. Imec is developing GeSe-based ovonic threshold switching (OTS) selector devices that fulfill the requirements imposed by high-density storage class memories, including high thermal and electrical stability, high current density and low off-state current.

Close to the DRAM side of the DRAM-NAND gap, high-density, high-speed MRAM offers an interesting, scalable alternative for the memory element. As this technology requires a selector as well, imec is working on a diode-based selector for this type of device. And finally, closer to the NAND Flash side, a ferroelectric memory based on hafnium-oxide (HfO) has gained interest. Compared to NAND-Flash, this emerging memory can operate at lower voltages and is much faster (with speeds down to 100ns). The easy cell structure can be fabricated with CMOS compatible processes and can be built in a 3D architecture.

3D NAND... and beyond?

Since its introduction several years ago, 3D NAND has become a mainstream storage technology because of its ability to significantly increase bit density. This is enabled by transitioning from 3 bits per cell to 4 bits per cell. And, instead of traditional x-y scaling in a horizontal plane, 3D NAND scales in the z direction by stacking multiple layers of NAND gates vertically. Today, stacking over 60 layers has become possible. However, increasing the numbers of layers challenges the deposition and etch

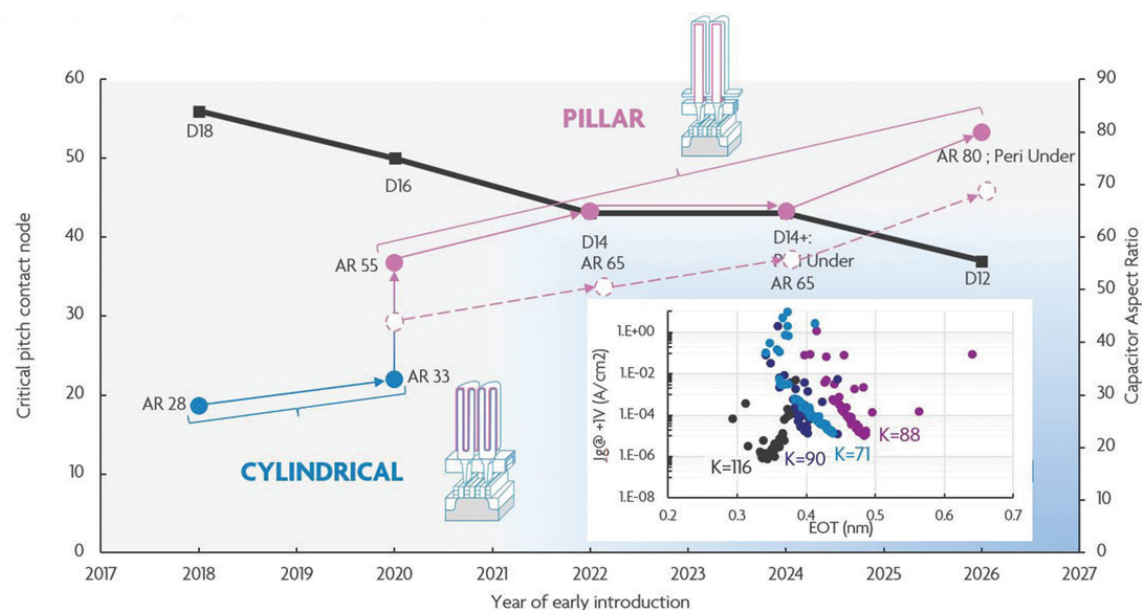


FIGURE 5. An imec view on the DRAM scaling roadmap; inset: impact of the one-pillar capacitor architecture

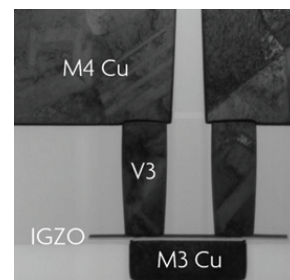


FIGURE 6. Novel oxide semiconductor based DRAM cell transistor.

processes. Also, the more layers that are stacked, the more stress evolves inside the layers, and this can cause a collapse of the 3D NAND pattern – a challenge that is tackled by imec. Imec also investigates alternative channel materials and processes – such as a silicon macaroni channel – that can overcome the limitations of the traditionally used poly-Si channels.

Despite all these advances, the density improvement of 3D NAND is expected to slow down and go below data growth rate soon. Therefore, the search for an emerging memory technology that is faster and cheaper than 3D NAND is ongoing. So far, there are no good candidates out there that can beat the 3D NAND density, especially because of the outstanding capability of 3D NAND Flash technology to integrate 3-4 bits per cell.

DNA storage: the holy grail of archival storage?

Imagine that we could store all data of the world in a container the size of a car, and store them for a very long time? That is exactly what DNA storage promises. DNA can be kept stable for millions of years – today, it is still possible to extract DNA from the woolly mammoth – guaranteeing long term retention. DNA as a medium for storage is also extremely dense and compact. Writing can be performed by encoding binary data onto strands of DNA through the process of DNA synthesis. The DNA strand can be built up with the base pairs representing a specific letter sequence, through a series of deprotection and protection reactions. As from the read side, there is an enormous technology push to sequence DNA faster and faster and at lower cost. Progress in DNA

sequencing has been amazing, even outpacing Moore's law. But researchers still have a long way to go before reasonable targets (1Gb/s) can be reached. To realize this, faster fluidics, faster chemical reactions and much higher parallelism are needed than what's possible today. At imec, we work towards faster write/read operation, and towards making DNA storage a cost-effective solution for long-term storage.

Towards a sustainable zettabyte era

It has become clear that the classical memory roadmap cannot handle the zettabyte world in terms of energy, density, speed and cost. As shown above, imec is working on several emerging memory and storage technologies that can largely improve on density, system performance, and, partly, speed. However, energy consumption remains the biggest challenge towards a truly sustainable zettabyte era. And this highlights the need to continue collaborating with academia and industry on reduced energy consumption for memory technologies.

Talking about sustainability brings another aspect of the zettabyte era to mind: recycling. To be able to process and store all the data, massive amounts of devices will be produced. The advent of emerging technologies will also bring in new materials, which today are hardly recycled. To enable a truly sustainable zettabyte era, the semiconductor industry should therefore also find ways to improve the recyclability of all these materials.

GOURI SANKARKAR is distinguished member technical staff emerging memories, and ARNAUD FURNEMONT is memory director at imec, Leuven, Belgium. ◀

There's still plenty of room at the bottom: Isotopically pure materials for when every atom counts

DR. PAUL STOCKMAN, Head of Market Development, Linde Electronics, Taipei, Taiwan

When different isotopes of atoms have significantly different properties, the ability to create isotopically pure materials becomes essential.

Nearly 60 years after Richard Feynman delivered his celebrated talk, which became the foundation for nanotechnology [1], many of the milestones he envisioned have been achieved and surpassed. In particular, he discussed computing devices with wires 10 to 100 atoms in width. Today we are reaching the smaller end of that range for high-volume FinFET and 10nm class DRAM chips, and device manufacturers are confidently laying the roadmaps for generations of conductors with single atom scales.

While device shrinkage has continued apace, it has not been without consequences. As chip circuit dimensions dip into the atomic range, bulk semiconductor properties which allowed for relatively simple scaling are breaking down and atomic-level physics are beginning to dominate.

At this scale, every atom counts. And when different isotopes of the needed atoms have significantly different properties, the ability to create isotopically pure materials (IPMs) becomes essential.

In this paper, we begin by discussing several examples of IPMs used in current high-volume electronics manufacturing: the physics at play and the materials selected. With the near future in focus, we then look at coming applications which may also require IPMs. Finally, we look at the current supply of IPM precursors, and how this needs to be developed in the future.

What is an isotope?

Isotopes are atoms of a particular element which have the same number of protons in their nucleus, but different numbers of neutrons (**FIGURE 1**). The number of protons determines which element the atom is: hydrogen has one proton, helium has two protons, and so on in the ordering used for the periodic table.

Different isotopes of the same element all have nearly exact chemical behavior – that is how they form and break molecular bonds in chemical reactions – but sometimes exhibit significantly different physical behavior. It is these differences in physical behavior which become important when electronics are made on the atomic scale.

For the purposes of engineering semiconductors, it is important to consider two different classes of isotopes.

- **Radioactive:** The nuclei of these isotopes are unstable, break apart into different and lighter elements, and often emit radiation in the form of alpha particles or light. The rate at which this

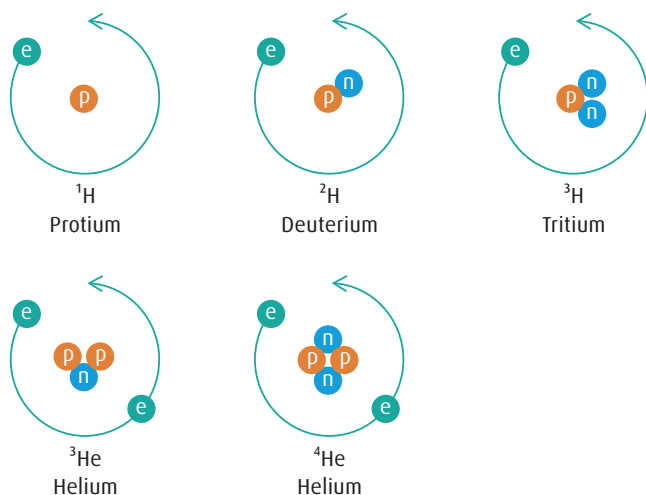


FIGURE 1. The stable isotopes of hydrogen and helium. The number of protons determines the element, and the number of neutrons determines the isotope. The superscript number is the sum of the protons and the neutrons.

happens can be quite fast – much less than a nanosecond – to longer than a billion years for half of the material to undergo decay. It is this type of isotope which was first historically observed and which we often first learn about. The element uranium has five different isotopes which naturally occur on earth, but all of these are radioactive.

- **Stable:** All other isotopes are termed stable. This means that they have not been observed to break apart, even once, when looking at bulk quantities of material with many billions of atoms. We do have evidence that some of these isotopes, termed primordial, are indeed stable over the span of knowable time, as we have evidence that they have formed and not decayed since the formation of the universe.

For the use of IPMs to engineer semiconductors, only stable isotopes are considered. Even for radioactive isotopes which decay slowly, the fact that current logic and memory chips have more than a billion transistors means that one or more circuits have the likelihood to be corrupted over the useful lifetime of the chip. In **FIGURE 2**, we show some common elements used in semiconductor manufacturing with their stable isotopes and naturally occurring abundances.

Current applications

There are already several electronics applications for IPMs, which have been used in high volume for more than a decade.

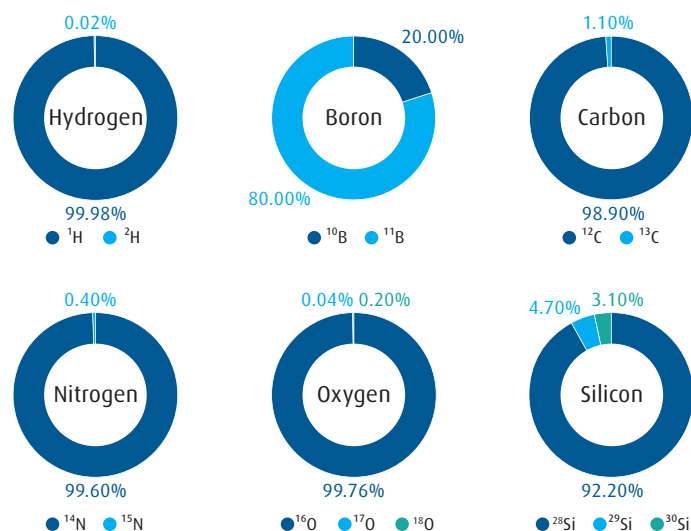


FIGURE 2. The natural abundance of stable isotopes for elements most relevant for electronics devices.

Deuterium ($\text{D}_2 = {}^2\text{H}_2$): Deuterium (D) is the second stable isotope of hydrogen, with one proton and one neutron. As a material, it is most commonly used in electronics manufacturing as the IPM precursor gas D_2 . Chemically, deuterium can be substituted directly for any reaction using normally abundant hydrogen. Deuterium is made by the electrolysis of D_2O , often called heavy water, which has been already enriched in the deuterium isotope.

Important physical property – mass:

For most elements, the difference in mass among their isotopes is only a few percent. However, for the lightest element hydrogen, there is a two-fold difference in mass. The chemical bond between a hydrogen atom and a heavier bulk material can be roughly approximated by the simple classical mechanics example of a weight at the end of the spring. When the weight is doubled, the force on the spring is also doubled (**FIGURE 3**).

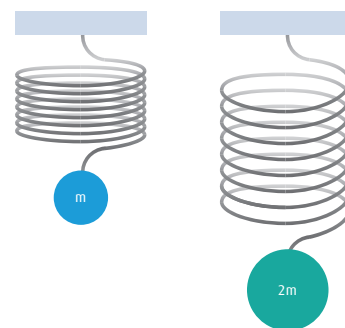


FIGURE 3. According to Hooke's Law, a spring will stretch twice as far when attached to a mass twice the original. When the balls and springs are bonds to naturally abundant hydrogen ${}^1\text{H}$ and deuterium ${}^2\text{H}$, they vibrate at different frequencies.

At the atomic level, quantum mechanics applies, and only certain amounts of energy, or quanta, can excite the spring. When deuterium ${}^2\text{H}$ is substituted for the much more abundant ${}^1\text{H}$, the amount of energy which can excite the spring changes by almost 50%.

Long-distance optical fibers: Optical fibers, like semiconductors, rely on silicon oxide as a primary material. The fiber acts as a waveguide to contain and transmit bursts of near-infrared lasers along the length of the fiber. The surface of the silicon oxide fiber is covered with hydroxide (oxygen-hydrogen), which is formed during the manufacture of the fiber. Unfortunately, the hydroxide chemical bonds absorb small amounts of the laser light with every single reflection against the surface, which in turn diminishes the signal. By substituting deuterium ${}^2\text{H}$ for normally abundant hydrogen ${}^1\text{H}$ on the surface hydroxide, the hydroxide molecular spring no longer absorbs light at the frequencies used for communication.

Hot carrier effect between gate and channel: As transistor sizes decrease, the local electric fields inside transistors increase. When the local field is high enough, it can generate free electrons with high kinetic energy,

known as hot carriers. Gate oxides are often annealed in hydrogen to reduce the deleterious effects of these hot carriers. But the hydride bonds themselves can become points of failure because the hot carriers have just the right amount of energy to excite and even break the hydride bonds. Just like in the optical fiber application, substituting deuterium ^2H for naturally abundant ^1H changes the energy of the bond and protects it against hot carrier damage. The lifetimes of the devices are extended by a factor of 50 to 100.

11-Boron trifluoride ($^{11}\text{BF}_3$): Boron has two naturally occurring stable isotopes ^{10}B at 20% and ^{11}B at 80%. Boron is used in electronics manufacturing as a dopant for silicon to modify its semiconducting properties, and is most commonly supplied as the gases boron trifluoride (BF_3) or diborane (B_2H_6). $^{11}\text{BF}_3$ is produced by the distillation of naturally abundant BF_3 , and can be converted to other boron compounds like B_2H_6 .

Important physical property – neutron capture:

The earth is continuously bombarded by high-energy cosmic radiation, which is produced primarily by events distant from our solar system. We are shielded from most of this radiation because it reacts with the molecules in our outer atmosphere. A by-product of this shielding mechanism are neutrons which constantly shower the earth's surface, but are not strong enough to pose any biological risk, usually passing through most materials without reaction. However, the nucleus of the ^{10}B atom is more than 1 million times likely to react with background neutrons versus other isotopes,

including ^{11}B . This results in splitting the ^{10}B atom into a ^7Li (lithium) atom and an alpha particle (helium nucleus) (**FIGURE 4**).

The smallest semiconductor gates now contain fewer than 100 dopant boron atoms. If even one of these is transmuted into a lithium atom, it can change the gate voltage and therefore the function of the transistor. Furthermore, the energetic alpha particle can cause additional damage. By using BF_3 which is depleted of ^{10}B below 0.1%, semiconductor manufacturers can greatly reduce the risk for component failure.

Initially, IPMs D_2 and $^{11}\text{BF}_3$ were used for making chips with the most critical value, like high performance computing processors, or remote operating environments like satellites and space vehicles. Now, as chip dimensions shrink to single nanometers and transistors multiply into the billions, these IPMs are increasingly being adopted into more high volume manufacturing processes.

Developing and future applications

As devices continue to scale to atomic dimensions and new device structures are developed to continue the progression of electronics advance, IPMs will play a greater role in a future where every atom counts. We present in this section a few of the nearest and most promising applications.

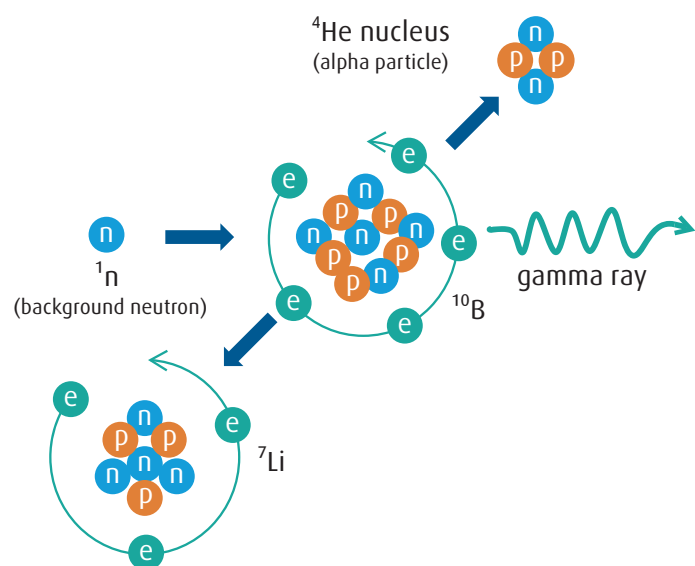


FIGURE 4. ^{10}B neutron capture. When a ^{10}B atom captures a background neutron, it breaks into a smaller ^7Li atom and emits alpha and gamma radiation.

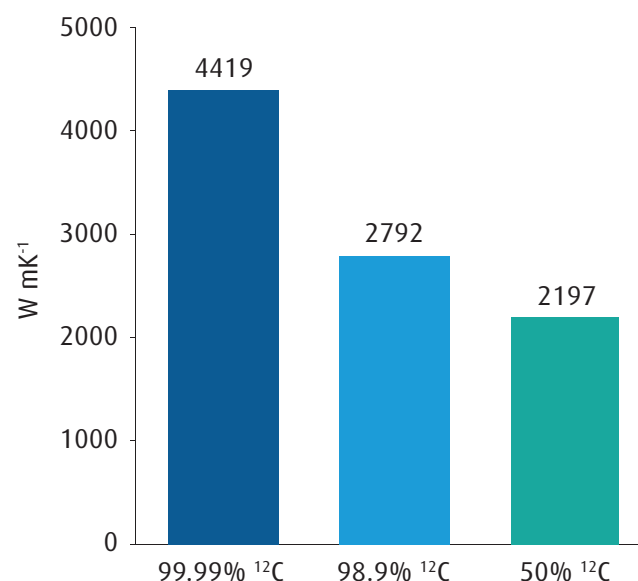


FIGURE 5. Thermal conductivity of graphene for different concentrations of ^{12}C content. 99.99% ^{12}C is achievable using commercial grade $^{12}\text{CH}_4$ methane; 98.6% ^{12}C is the value of naturally abundant carbon on earth; 50% ^{12}C is an artificial mixture to demonstrate the trend. Source: Thermal conductivity of isotopically modified graphene. S Chen et al., Nature Materials volume 11, pages 203–207 (2012).

Important physical property – thermal conductivity:

Thermal conductivity is the property of materials to transfer heat, which is especially important in semiconductor chips where localized transistor temperatures can exceed 150 C and can affect the chip performance. Materials like carbon (either diamond or graphene) and copper have relatively high thermal conductivity; silicon and silicon nitride medium and oxides like silicon oxide and aluminum oxide are much lower. However, with all of the other design constraints and requirements, semiconductor engineers seldom choose materials to optimize thermal conductivity.

When viewed in classical mechanics, thermal conductivity is like a large set of balls and springs. In an atomically pure material like diamond or silicon, all the springs are the same, and all the balls are nearly the same – the differences being the different masses of the naturally occurring isotopes. When IPMs are used to make the material, all the balls are now exactly the same, there is less disorder, and heat is transferred through the matrix of balls and springs more efficiently. This has been demonstrated in graphene to improve the thermal conductivity at room temperature by 60% [FIGURE 5], and other studies have shown a similar magnitude improvement in silicon.

Silicon epilayers: Epitaxially grown silicon is often the starting substrate for CMOS manufacturing. This may be especially beneficial for HD-SOI applications where trade-offs of substrate cost vs processing cost and device performance are already part of the value equation.

Sub 3nm 2D graphene FETs: Graphene is a much-discussed material being considered for sub-3nm devices as the successor technology for logic circuits after the FinFET era. Isotopically pure graphene could reduce localized heating at the source.

Important physical property – nuclear spin: Each atomic nucleus has a quantum mechanical property associated with it called nuclear spin. Because it is a quantum mechanical property, nuclear spin is measured in discrete amounts, and in this case half-integer numbers. Nuclear spin is determined by the number of protons and neutrons. Since different isotopes have different numbers of neutrons, they also have different nuclear spin. An atomically and isotopically pure material will have atoms with all the same spin.

Qubits and quantum computing: Much research has been published recently about quantum computing as the successor to transistor-based processors. IBM and Intel, among others, have made demonstration devices, albeit not large enough for practical applications yet. The most

promising near-term realization of quantum computing uses qubits – atomic two-state components—which store and transmit information via electron spin, which is a property analogous to nuclear spin, and can be affected by nuclear spin. Many of these early devices have been made with isotopically pure diamond (carbon) or silicon matrices to avoid disorder from having multiple values of nuclear spin in these devices.

12-Methane ($^{12}\text{CH}_4$): Carbon is predominantly ^{12}C , with about 1.1% ^{13}C in natural abundance. A large demand already exists for ^{13}C chemicals, primarily used as markers in studying chemical and biological reactions. ^{13}C is produced primarily by the distillation of carbon monoxide (CO), and then chemically converted to other carbon-containing precursors like methane (CH_4). At the same time ^{13}C is produced, a large amount of ^{13}C -depleted ^{12}CO is produced, which serves as a less expensive feedstock for applications which require IPMs made from carbon.

28-Silane ($^{28}\text{SiH}_4$) and other silicon precursors: There are no such large applications driving the production of isotopically pure silicon precursors yet. Currently, research quantities of silicon tetrafluoride (SiF_4) are produced primarily by using gas centrifuges, and then converted into silicon precursors like silicon tetrachloride (SiCl_4), trichlorosilane (SiHCl_3), dichlorosilane (SiH_2Cl_2), and silane (SiH_4). Distillation of one or more of these materials would be a less expensive option for larger-scale production.

Other IPM precursors already exist for oxygen and nitrogen compounds, and are relatively inexpensive because they are also by-products from the production of chemical markers for less abundant isotopes. Aluminum and phosphorous only have one stable isotope, and so all compounds produced with these are isotopically pure in these elements.

Production methods

Production of IPMs is challenging because of the limited differences in physical and chemical properties normally used to separate and purify materials, and because of the low concentration of some of the desired isotopes. A number of creative approaches have been applied, and often the methods are repeated to obtain the desired enrichment and purity of the IPM. We give a description here of the two methods used today in the production of IPMs relevant to the electronics industry. Importantly, all of these require gas-phase starting materials in order to enhance the physical differences that do exist among the isotopes.

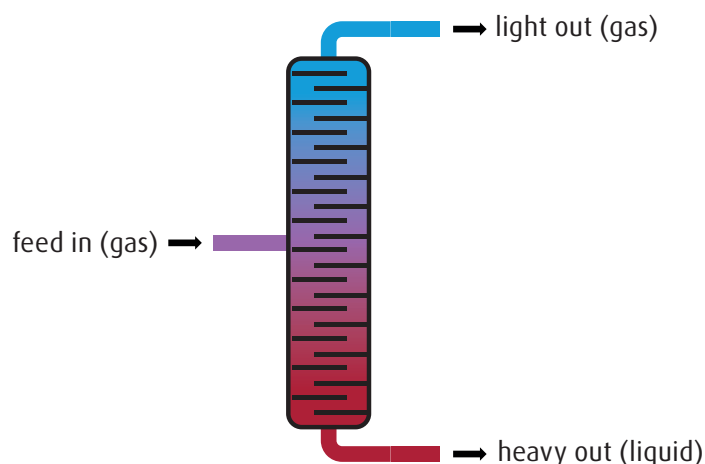


FIGURE 6. Isotope separation. Distillation. Gas is fed into the middle of the distillation column. By the process of condensation and boiling on many plates, the lighter isotope is separated as a gas leaving the top of the column, and the heavier isotope leaves as a liquid at the bottom.

Distillation: The more familiar of the two is distillation, which relies on differing boiling points of materials to separate them into lower boiling point fractions called *lights* (which often are lighter in mass) and higher boiling point fractions which are called *heavies*. Because the boiling point differences are a fraction of a degree, very long distillation columns are used. Distillation is best used with lighter compounds, and is the preferred method for producing D_2O , $^{11}BF_3$, $^{12}CO/^{13}CO$, and isotopes of nitrogen and oxygen (**FIGURE 6**).

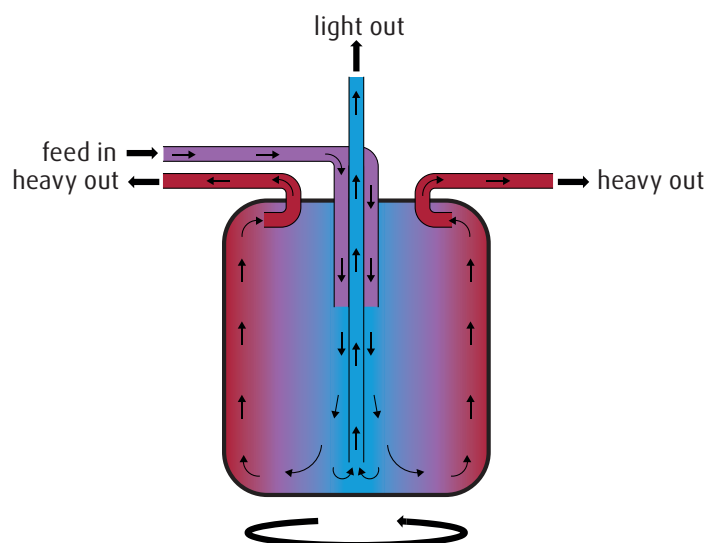


FIGURE 7. In this isotope separation centrifuge, gas is fed into the center of the centrifuge, which is spinning at a very high rate of 100,000 rpm. The heavier isotope is thrown to the sides, while the lighter isotope remains in the center.

Centrifuge: After a number of other methods were tried, gas centrifuges were the method ultimately chosen to scale the separation of the ^{235}U uranium isotope in uranium hexafluoride (UF_6) gas used for the first atomic devices during the Manhattan Project, and remains the preferred method of obtaining this useful isotope and other isotopes of heavier elements. The gas is spun at very high speeds – around 100,000 rpm – and the higher mass isotopes tend toward the outer regions of the centrifuge. Many gas centrifuges are linked in arrays to achieve the desired level of enrichment. Currently, useful quantities of $^{28}SiF_4$ and $^{30}SiF_4$ are produced with this method (**FIGURE 7**).

Conclusion

In the hundred-year anniversary of Richard Feynman's birth, we are still finding plenty of room at the bottom. But as we go further down, we must look more carefully at what is there. Increasingly, we are seeing that individual atoms hold the properties that are important today and which will support the developments of a not-too-distant tomorrow. IPMs are an important part of creating that reality.

Linde Electronics is the leader in the production of IPMs which are important to electronics today, and holds the technology to produce the IPMs which will support the development of tomorrow's devices. Linde has made recent investments for the production of deuterium (D_2) and $^{11}BF_3$ to satisfy global electronics demand, and has a long history in the production, purification, and chemical synthesis of stable isotopes relevant to semiconductor manufacturing.

Dr. PAUL STOCKMAN is Head of Market Development, Linde Electronics, Taipei, Taiwan.

Reference

1. Feynman, Richard P. (1960) There's Plenty of Room at the Bottom. *Engineering and Science*, 23 (5). pp. 22-36. ◀

AI chips: Challenges and opportunities

PETE SINGER, Editor-in-Chief

To get to the next level in performance/Watt, innovations being researched at the AI chip level include low precision, analog and resistive computing.

The exploding use of Artificial Intelligence (AI) is ushering in a new era for semiconductor devices that will bring many new opportunities but also many challenges. Speaking at the AI Design Forum hosted by Applied Materials and SEMI during SEMICON West in July, Dr. John E. Kelly, III, Senior Vice President, Cognitive Solutions and IBM Research, talked about how AI will dramatically change the world. “This is an era of computing which is at a scale that will dwarf the previous era, in ways that will change all of our businesses and all of our industries, and all of our lives,” he said. “This is the era that’s going to power our semiconductor industry forward. The number of opportunities is enormous.”

Also speaking at the event, Gary Dickerson, CEO of Applied Materials, said AI “needs innovation in the edge and in the cloud, in generating data on the edge, storing the data, and processing that data to unlock the value. At the same time Moore’s Law is slowing.” This creates the “perfect opportunity,” he said.

Ajit Manocha, President and CEO of SEMI, calls it a “rebirth” of the semiconductor industry. “Artificial Intelligence is changing everything – and bringing semiconductors back into the deserved spotlight,” he notes in a recent article. “AI’s potential market of hundreds of zettabytes and trillions of dollars relies on new semiconductor architectures and compute platforms. Making these AI semiconductor engines will require a wildly innovative range of new materials, equipment, and design methodologies.”

“Hardware is becoming sexy again,” said Dickerson. “In the last 18 months there’s been more money going into chip start ups than the previous 18 years.” In addition to AI chips from traditional IC companies such as Intel and Qualcomm, more than 45 start-ups are working to develop new AI chips, with VC investments of more than \$1.5B — at least five of them have raised more

than \$100 million from investors. Tech giants such as Google, Facebook, Microsoft, Amazon, Baidu and Alibaba are also developing AI chips.

Dickerson said having the winning AI chip 12 months ahead of anyone else could be a \$100 billion opportunity. “What we’re driving inside of Applied Materials is speed and time to market. What is one month worth? What is one minute worth?”

IBM’s Kelly said there’s \$2 trillion of decision support opportunity for artificial intelligence on top of the existing \$1.5-2 billion information technology industry. “Literally every industry in the world is going to be impacted and transformed by this,” he said.

AI needed to analyze unstructured data

Speaking at an Applied Materials event late last year during the International Electron Devices Meeting, Dr. Jeff Welser, Vice President and Director of IBM Research’s – Almaden lab, said the explosion in AI is being driven by the need to process vast amounts of unstructured data, noting that in just two days, we now generate as much data as was generated in total through 2003. “Somewhere around 2020, the estimate is maybe 50 zettabytes of data being produced. That’s 21 zeros,” he said.

Welser — who will be delivering the keynote talk at The ConFab 2019 in May — noted that 80% of all data is unstructured and growing 15 times the rate of structured data. “If you look at the growth, it’s really in a whole different type of data. Voice data, social media data, which includes a lot of images, videos, audio and text, but very unstructured text,” he said. And then there’s data from IoT-connected sensors.

Custom hardware for AI is not new. “Even as early as the ‘90s, they were starting to play around with ASICs and FPGAs, trying to find ways to do this better,” Welser said.

Google's Tensor Processing Unit (TPU), introduced in 2016, for example, is a custom ASIC chip built specifically for machine learning applications, allowing the chip to be more tolerant of reduced computational precision, which means it requires fewer transistors per operation.

It really was when the GPUs appeared in the 2008-2009 time period when people realized that in addition to the intended application – graphics processing – they were really good for doing the kind of math needed for neural nets. “Since then, we’ve seen a whole bunch of different architectures coming out to try to continue to improve our ability to run the neural net for training and for inferencing,” he said.

AI works by first “training” a neural network where weights are changed based on the output, followed by an “inferencing” aspect where the weights are fixed. This may mean two different kinds of chips are needed. “If you weren’t trying to do learning on it, you could potentially get something that’s much lower power, much faster, much more efficient when taking an already trained neural net and running it for whatever application. That turns out to be important in terms of where we see hardware going,” he said.

The problem with present day technology – whether it’s CPUs, GPUs, ASICs or FPGAs — is that there is still a huge gap between what processing power is required and what’s available now. “We have a 1,000x gap in performance per watt that we have to close,” said Applied Materials’ Dickerson.

There’s a need to reduce the amount of power used in AI processors not only at data centers, but for mobile applications such as automotive and security where decisions need to be made in real time versus in the cloud. This also could lead to a need for different kinds of AI chips.

An interesting case in point: IBM’s world-leading Summit supercomputer, employs 9,216 IBM processors boosted by 27,648 Nvidia GPUs – and takes a room the size of two tennis courts and as much power as a small town!

New approaches

To get to the next level in performance/Watt, innovations being researched at the AI chip level include:

- low precision computing
- analog computing
- resistive computing

In one study, IBM artificially reduced the precision in a neural net and the results were surprising. “We found we could get down the floating point to 14 bit, and we really were getting exactly the same precision as you could with 16 bit or 32 bit or 64 bit,” Welser said. “It didn’t really matter at that point.”

This means that some parts of the neural net could be high precision and some parts that are low precision. “There’s a lot of tradeoffs you can make there, that could get you lower power or higher performance for that power, by giving up precision,” Welser said.

Old-school analog computing has even lower precision but may be well suited to AI. “Analog computing was extremely efficient at the time, it’s just you can’t control the errors or scale it in any way that makes sense if you’re trying to do high precision floating point,” Welser said. “But if what you really want is the ability to have a variable connection, say to neurons, then perhaps you could actually use an analog device.”

Resistive computing is a twist on analog computing that has the added advantage of eliminating the bottleneck between memory and compute. Welser said to think of it as layers of neurons, and the connections between those neurons would be an analog resistive memory. “By changing the level of that resistive memory, the amount of current that flows between one neuron and the next would be varied automatically. The next neuron down would decide how it’s going to fire based on the amount of current that flowed into it.

IBM experimented with phase change memory for this application. “Obviously phase change memory can go to a low resistance or a high resistance (i.e., a 1 or a 0) but there is no reason you can’t take it somewhere in between, and that’s exactly what we would want to take advantage of here,” Welser said.

“There is hope for taking analog devices and using them to actually be some of the elements and getting rid of the bottleneck for the memory as well as getting away from the precision/power that goes on with trying to get to high precision for those connections,” he added.

A successful resistive analog memory ultimately winds up being a materials challenge. “We’d like to have like a thousand levels for the storage capacity, and we’d like to have a very nice symmetry in turning it off and on, which is not something you’d normally think about,” Welser said. “One of the challenges for the industry is to think about how you can get materials that fit these needs better than just a straight memory of one bit on or off.”◆

New thinking required for machine learning

DAVE LAMMERS, Contributing Editor

Speakers argue the semiconductor community thus far has not been doing enough to enable machine intelligence.

Judging by the presentations at the 2018 Symposium on VLSI Technology, held in Honolulu this summer, the semiconductor industry has a challenge ahead of it: how to develop the special low-power hardware needed to support artificial intelligence-enabled networks.

To meet society's needs for low-power-consumption machine learning (ML), "we do need to turn our attention to this new type of computing," said Naveen Verma, an associate professor of electrical engineering at Princeton University.

While introducing intelligence into engineering systems has been what the semiconductor industry has been all about, Verma said machine learning represents a "quite distinct" inflection point. Accustomed as it is to fast-growing applications, machine learning is on a growth trajectory that Verma said is "unprecedented in our own industry" as ML algorithms have started to outperform human capabilities in a wide variety of fields.

Faster GPUs driven by Moore's Law, and combining chips in packages by means of heterogeneous computing, "won't be enough as we proceed into the future. I would suggest we need to do something more, get engaged more deeply, affecting things done at all levels."

Naresh Shanbhag, a professor at the University of Illinois at Urbana-Champaign, sounded a similar appeal at the VLSI symposium's day-long workshop on machine learning. The semiconductor industry has taken a "back seat to the systems and algorithm researchers who are driving the AI revolution today," he said. Addressing several hundred device and circuit researchers, Shanbhag said their contributions to the AI revolution have been hampered by a self-limiting mindset, based on "our traditional role as component providers."

Until last year, Shanbhag served as the director of a multi-university research effort, Systems on Nanoscale Information fabriCs (SONIC, www.sonic-center.org), which pursued new forms of low-power compute networks, including work on fault-tolerant computing. At the VLSI symposium he spoke on Deep In-Memory Architectures and other non-traditional approaches.

"Traditional solutions are running out of steam," he said, noting the slowdown in scaling and the "memory wall" in traditional von Neumann architectures that contributes to high power consumption. "We need to develop a systems-to-devices perspective in order to be a player in the world of AI," he said.

Boris Murmann, an associate professor in the Department of Electrical Engineering at Stanford University, described a low-power approach based on mixed signal-based processing, which can be tightly coupled to the sensor front-end of small-form-factor applications, such as an IoT edge camera or microphone.

"What if energy is the prime currency, how far can I push down the power consumption?" Murmann asked. By coupling analog-based computing to small-scale software macros, an edge camera could be awakened by a face-recognition algorithm. The "wake-up triggers" could alert more-powerful companion algorithms, in some cases sending data to the cloud.

Showing a test chip of mixed-signal processing circuits, Murmann said the Stanford effort "brings us a little bit closer to the physical medium here. We want to exploit mixed-signal techniques to reduce the data volume, keeping it close to its source." In addition, mixed-signal computing could help lower energy consumption in wearables, or in convolutional neural networks (CNNs) in edge IoT devices.

In a remark that coincided with others' views, Murmann said "falling back to non-deep learning techniques can be advantageous for basic classification tasks," such as wake-up-type alerts. "There exist many examples of the benefits of analog processing in non-deep learning algorithms," he said.

That theme – when deep learning competes with less power-hungry techniques – was taken up by Vivienne Sze, an associate professor at the Massachusetts Institute of Technology. By good fortune, Sze recently had two graduate students who designed similar facial recognition chips, one based on the Histograms of Oriented Gradients (HOG) method of feature recognition, and the other using the MIT-developed Eyeriss accelerator for CNNs (eyeriss.mit.edu). Both chips were implemented in the same foundry technology, with similar logic and memory densities, and put to work on facial recognition (**FIGURE 1**).

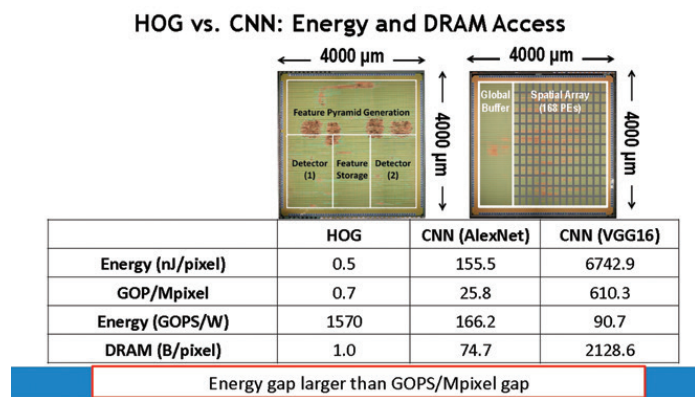


FIGURE 1. Two image processing chips created at M.I.T. resulted in sharply different energy consumption levels. Machine learning approaches, such as CNNs, are flexible, but often not as efficient as more hard-wired solutions. (Source: 2018 VLSI Symposium).

Calling it a "good controlled experiment," Sze described the energy consumption versus accuracy measurements, concluding that the Eyeriss machine-learning chip was twice as accurate on the AlexNet benchmark. However, that doubling in accuracy came at the price of a 300-times multiplier in energy, increasing to a 10,000-times energy penalty in some cases, as measured in nanojoules per pixel.

"The energy gap was much larger than the throughput gap," Sze said. The Eyeriss CNNs require more energy because of the programmability factor, with weights of eight bits per pixel. "The question becomes are you willing to give up a 300x increase in energy, or even 10,000x, to get a 2X increase in accuracy? Are you willing to sacrifice that much battery life?"

"The main point -- and it is really important -- is that CNNs are not always the best solution. Some hand-crafted features perform better," Sze said.

Two European consortia, CEA-Leti and Imec, were well represented at the VLSI symposium.

Denis Dutoit, a researcher at France's CEA Tech center, described a deep learning core, PNeuro, designed for neural network processing chains.

The solution supports traditional image processing chains, such as filtering, without external image processing. The modular SIMD architecture can be sized to fit the best area/performance per application.

Dutoit said the energy consumption was much less than that of traditional cores from ARM and Nvidia on a benchmark application, recognizing faces from a database of 18,000 images at a recognition rate of 97 percent.

GPUs vs custom accelerators

The sharp uptake of AI in image and voice recognition, navigation systems, and digital assistants has come in part because the training cycles could be completed efficiently on massively parallel architectures, i.e., GPUs, said Bill Dally, chief scientist at Nvidia Inc. Alternatives to GPUs and CPUs are being developed that are faster, but less flexible. Dally conceded that creating a task-specific processor might result in a 20 percent performance gain, compared with a GPU or Transaction Processing Unit (TPU). However, "you would lose flexibility if the algorithm changes. It's a continuum: (with GPUs) you give up a little efficiency while maximizing flexibility," Dally said, predicting that "AI will dominate loads going forward."

Joe Macri, a vice president at AMD, said that modern processors have high-speed interfaces with "lots of coherancy," allowing dedicated processors and CPUs/GPUs to use shared memory. "It is not a question of an accelerator or a CPU. It's both."

Whether it is reconfigurable architectures, hard-wired circuits, and others, participants at the VLSI symposium agreed that AI is set to change lives around the globe. Macri pointed out that only 20 years ago, few people carried phones. Now, no one would even think of going out with their smart phone – it has become more important than carrying a billfold or purse, he noted. Twenty years from now, machine learning will be embedded into phones, homes, and factories, changing lives in ways few of us can foresee. ◀▶

Dynamic Fault Detection: Utilizing AI and IoT to revolutionize manufacturing

TOM HO and **STEWART CHALMERS**, Tom Ho Stewart Chalmers, BisTEL, Santa Clara, CA

A new approach in Fault Detection and Classification (FDC) allows engineers to uncover issues more thoroughly and accurately by taking advantage of full sensor traces.

Traditional FDC systems collect data from production equipment, summarize it, and compare it to control limits that were previously set up by engineers. Software alarms are triggered when any of the summarized data fall outside of the control limits. While this method has been effective and widely deployed, it does create a few challenges for the engineers:

- The use of summary data means that (1) subtle changes in the process may not be noticed and (2) the unmonitored section of the process will be overlooked by a typical FDC system. These subtle changes or the missed anomalies in unmonitored section may result in critical problems.
- Modeling control limits for fault detection is a manual process, prone to human error and process drift. With hundreds of thousands of sensors in a complex manufacturing process, the task of modeling control limits is extremely time consuming and requires a deep understanding of the particular manufacturing process on the part of the engineer. Non-optimized control limits result in misdetection: false alarms or missed alarms.
- As equipment ages, processes change. Meticulously set control limit ranges must be adjusted, requiring engineers to constantly monitor equipment and sensor data to avoid false alarms or missed real alarm.

Full sensor trace detection

A new approach, Dynamic Fault Detection (DFD) was developed to address the shortcomings of traditional FDC systems and save both production time and engineer time. DFD takes advantage of the full trace from each and every sensor to detect any issues during a manufacturing process. By analyzing each trace in its entirety, and running them through intelligent software, the system is able to comprehensively identify potential

issues and errors as they occur. As the Adaptive Intelligence behind Dynamic Fault Detection learns each unique production environment, it will be able to identify process anomalies in real time without the need for manual adjustment from engineers. Great savings can be realized by early detection, increased engineer productivity, and containment of malfunctions.

DFD's strength is its ability to analyze full trace data. As shown in **FIGURE 1**, there are many subtle details on a trace, such as spikes, shifts, and ramp rate changes, which are typically ignored or go undetected by a traditional FDC system, because they only examine a segment of the trace- summary data. By analyzing the full trace using DFD, these details can easily be identified to provide a more thorough analysis than ever before.

Dynamic referencing

Unlike traditional FDC deployments, DFD does not require control limit modeling. The novel solution adapts machine learning techniques to take advantage of neighboring traces as references, so control limits are dynamically defined in real time. Not only does this substantially reduce set up and deployment time of a fault detection system, it also eliminates the need for an engineer to continuously maintain the model. Since the analysis is done in real time, the model evolves and adapts to any process shifts as new reference traces are added. DFD has multiple reference configurations available for engineers to choose from to fine tune detection

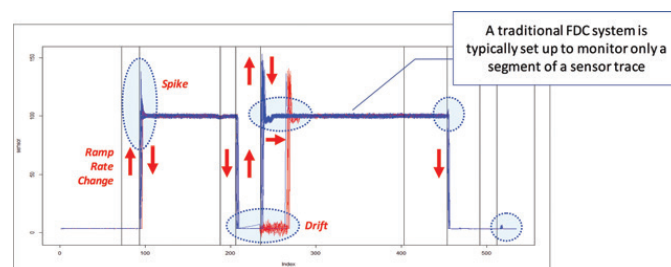


FIGURE 1.

accuracy. For example, DFD can 1) use traces within a wafer lot as reference, 2) use traces from the last N wafers as reference, 3) use “golden” traces as reference, or 4) a combination of the above. As more sensors are added to the Internet of Things network of a production plant, DFD can integrate their data into its decision-making process.

Optimized alarming

Thousands of process alarms inundate engineers each day, only a small percentage of which are valid. In today's FDC systems, one of the main causes for false alarms is improperly configured Statistical Process Control (SPC) limits. Also, typical FDC may generate one alarm for each limit violation resulting in many alarms for each wafer process. DFD implementations require no control limits, greatly reducing the potential for false alarms. In addition, DFD is designed to only issue one alarm per wafer, further streamlining the alarming system and providing better focus for the engineers.

Dynamic fault detection use cases

The following examples illustrate actual use cases to show the benefits of utilizing DFD for fault detection.

Use case #1 – End Point Abnormal Etching

In this example, both the upper and lower control limits in SPC were not set at the optimum levels, preventing the traditional FDC system from detecting several abnormally etched wafers (**FIGURE 2**). No SPC alarms were issued to notify the engineer.

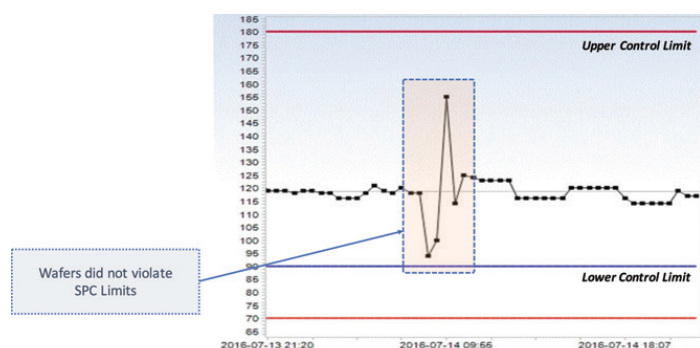


FIGURE 2.

On the other hand, DFD full trace comparison easily detects the abnormality by comparing to neighboring traces (**FIGURE 3**). This was accomplished without having to set up any control limits. show the benefits of utilizing DFD for fault detection.

Use case #2 – Resist Bake Plate Temperature

The SPC chart in **FIGURE 4** clearly shows that the Resist bake plate temperature pattern changed

significantly; however, since the temperature range during the process never exceeded the control limits, SPC did not issue any alarms.

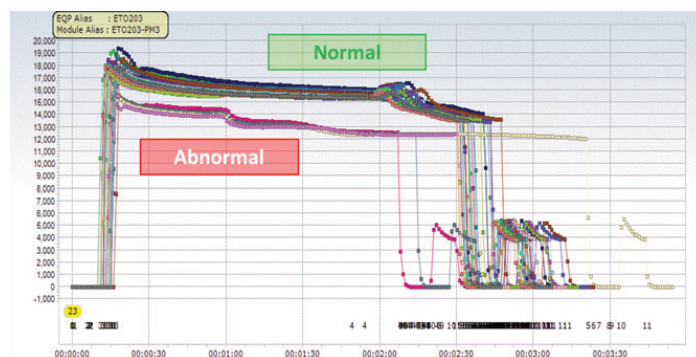


FIGURE 3.

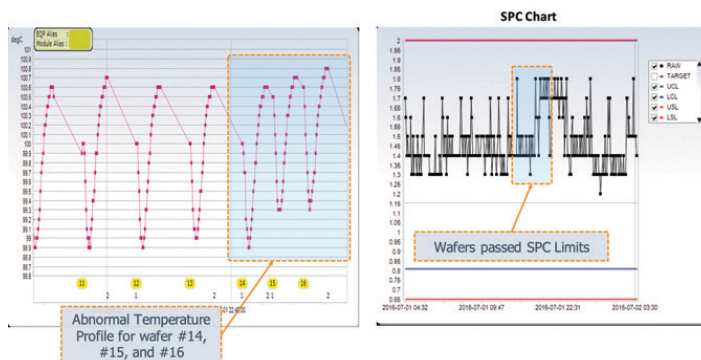


FIGURE 4.

When the same parameter was analyzed using DFD, the temperature profile abnormality was easily identified, and the software notified an engineer (**FIGURE 5**).

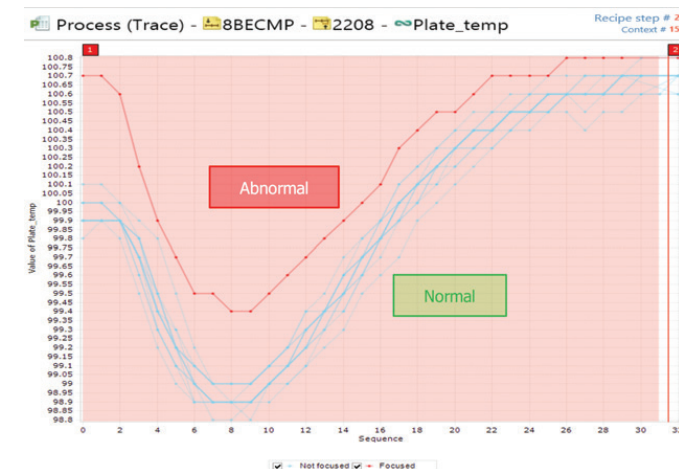


FIGURE 5.

Use case #3 – Full Trace Coverage

Engineers select only a segment of sensor trace data to monitor because setting up SPC limits is so arduous. In this specific case, the SPC system was set up to monitor only the He_Flow parameter in recipe step 3 and step 4. Since no unusual events occurred during those steps in the process, no SPC alarms were triggered.

However, in that same production run, a DFD alarm was issued for one of the wafers. Upon examination of the trace summary chart shown in **FIGURE 6**, it is clear that while the parameter behaved normally during recipe step 3 and step 4, there was a noticeable issue from one of the wafers during recipe step 1 and step 2. The trace in red represents the offending trace versus the rest of the (normal) population in blue. DFD full trace analysis caught the abnormality.



FIGURE 6.

Use case #4 – DFD Alarm Accuracy

When setting up SPC limits in a conventional FDC system, the method of calculation taken by an engineer can yield vastly different results. In this example, the engineer used multiple SPC approaches to monitor parameter *Match_LoadCap* in an etcher. When the control limits were set using *Standard Deviation* (**FIGURE 7**), a large number of false alarms were triggered. On the other hand, zero alarms were triggered using the *Mean* approach (**FIGURE 8**).

Using DFD full trace detection eliminates the discrepancy between calculation methods. In the above example, DFD was able to identify an issue with one of the wafers in recipe step 3 and trigger only one alarm.

Dynamic fault detection scope of use

DFD is designed to be used in production environments of many types, ranging from semiconductor

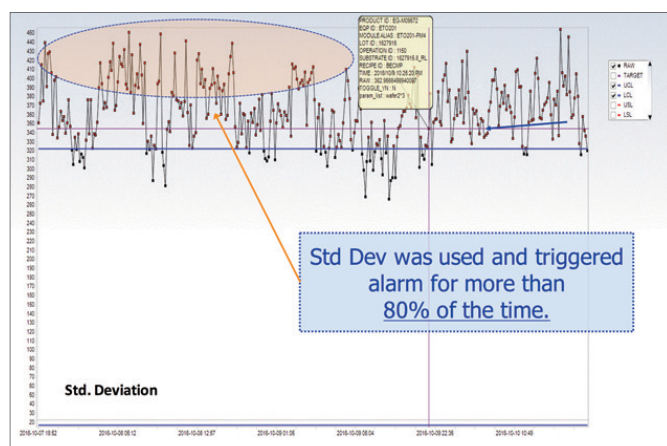


FIGURE 7.

www.solid-state.com

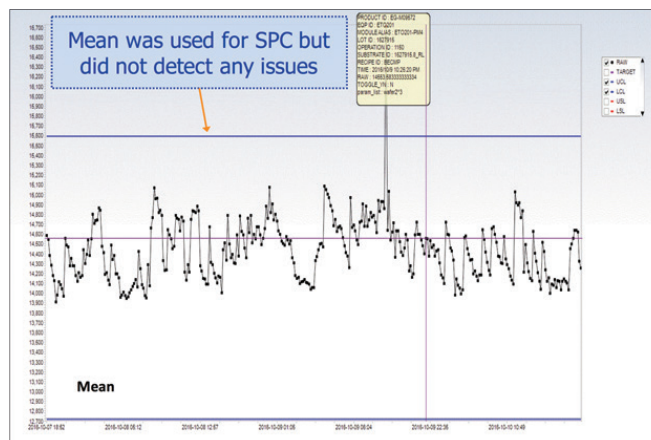


FIGURE 8.

manufacturing to automotive plants and everything in between. As long as the manufacturing equipment being monitored generates systematic and consistent trace patterns, such as gas flow, temperature, pressure, power etc., proper referencing can be established by the Adaptive Intelligence (AI) to identify abnormalities. Sensor traces from Process of Record (POR) runs may be used as starting references.

Conclusion

The DFD solution reduces risk in manufacturing by protecting against events that impact yield. It also provides engineers with an innovative new tool that addresses several limitations of today's traditional FDC systems. As shown in **TABLE 1**, the solution greatly reduces the time required for deployment and maintenance, while providing a more thorough and accurate detection of issues.

	FDC (Per Recipe/Tool)	DFD (Per Recipe/Tool)
FDC model creation	1 – 2 weeks	< 1 day
FDC model validation and fine tuning	2 – 3 weeks	< 1 week
Model Maintenance	Ongoing	Minimal
Typical Alarm Rate	100-500/chamber-day	< 50/chamber-day
% Coverage of Number of Sensors	50-60%	100% as default
Trace Segment Coverage	20-40%	100%
Adaptive to Systematic Behavior Changes	No	Yes

TABLE 1.

TOM HO is President of BISTel America where he leads global product engineer and development efforts for BISTel. tomho@bistel.com. STEWART CHALMERS is President & CEO of Hill + Kincaid, a technical marketing firm. stewart@hillandkincaid.com. ◊



senior
Metal Bellows

PRIME MOVER™ - All Metal Pneumatic Actuators

- No elastomers for smooth continuous motion across the vacuum boundary
- Integrated position sensing with automatic temperature compensation
- Precision guiding for accuracy and repeatability
- Individually tuned for synchronous motion
- Very wide operating temperature range
- Million+ Lifecycle

www.metalbellows.com



ULVAC

Non-volatile Memory Solutions

ULVAC Technologies deposition and etching equipment provides fabrication solutions for non-volatile memory technologies. Whether it is MRAM, PCRAM, ReRAM, FeRAM, CBRAM or STT-MRAM, ULVAC has unique and novel solutions ideal for fabrication of these memory technologies.

Contact: sales@us.ulvac.com or 978-686-7550.

www.ulvac.com



YES-ÉcoClean

- Automated Plasma Resist Strip/Descum System
- 2x faster, 1/2 the capital cost, and lower CoO
- Small footprint with Single Chamber
- Flexible System from Descum to Strip – 100 to 100,000 Å/min
- No defects or damage due to ICP Downstream Plasma
- R&D to high volume production with minimal downtime
- Eco-friendly “Green” process

1-888-YES-3637

www.yieldengineering.com



Magazine

Newsletter

Website

Solid State Technology

For over 50 years, Solid State Technology has been the leading independent media resource, covering:

- Semiconductors
- Advanced Packaging
- MEMS
- Displays
- LEDs

Request a FREE subscription to our magazine and e-Newsletters today at www.solid-state.com/subscribe

www.solid-state.com

ad index

Advertiser	Pg
Conax	11
ConFab	9
Levitronix	C4
MEPTEC	7
Nikon Precision	13
Park Systems	C2
SEMI	3
Senior Metal Bellows	30
Ulvac	30
Y.E.S.	C3, 30

Solid State TECHNOLOGY

EXECUTIVE OFFICES

Extension Media 1786 18th Street, San Francisco, CA 94107-2343.

ADVERTISING

Sales Manager
Kerry Hoffman
1786 18th St.
San Francisco, CA 94107-2343
Tel: 978.580.4205
khoffman@extensionmedia.com

North America
Kerry Hoffman
Tel: 978.580.4205
khoffman@extensionmedia.com

Germany, Austria, E. Switzerland & E. Europe
Holger Gerisch
Tel: +49.0.8856.8020228
holgerg@pennwell.com

China, Hong Kong
Adonis Mak
Tel: +852.90182962
adonism@actintl.com.hk

Taiwan
Diana Wei
Tel: +886.2.23965128 ext: 270
diana@arco.com.tw

Rest of World
Kerry Hoffman
Tel: 978.580.4205
khoffman@extensionmedia.com

Webcasts
Jenna Johnson
Tel: 612.598.3446
jjohnson@extensionmedia.com

The ConFab
Kerry Hoffman
Tel: 978.580.4205
khoffman@extensionmedia.com

Sensors in the new age of the car

RICHARD DIXON, Senior Principal Analyst, Sensors, IHS Markit

Sensors are inextricably linked to the future requirements of partially and fully autonomous vehicles. From highly granular dead-reckoning subsystems that rely on industrial-strength gyroscopes for superior navigation to more intelligent and personalized cockpits featuring intuitive human machine interfaces (HMIs) and smart seats, new generations of partially and fully autonomous cars will use sensors to enable dramatically better customer experiences.

Dead reckoning, or, where am I, exactly?

Dead reckoning is the process of calculating one's current position by using a previously determined position, and advancing that position based upon known speeds over a time slice. As a highly useful process, dead reckoning is the basis for inertial navigation systems in aerospace navigation and missile guidance, not to mention your smartphone.

Today's best-in-class MEMS gyroscopes can offer 30-50 cm resolution (this is the yaw rate drift) over a distance of 200 m—a typical tunnel length where a GPS signal is lost. For semi-autonomous (L3) or autonomous (L4, L5), the locational accuracy is well below 10 centimeters; that's an accuracy usually reserved for high-end industrial or aerospace gyroscopes with a raw bias instability ranging from 1°/h and down to 0.01°/h. These heavy-duty gyros command prices from \$100s up to \$1000s (**FIGURE 1**).

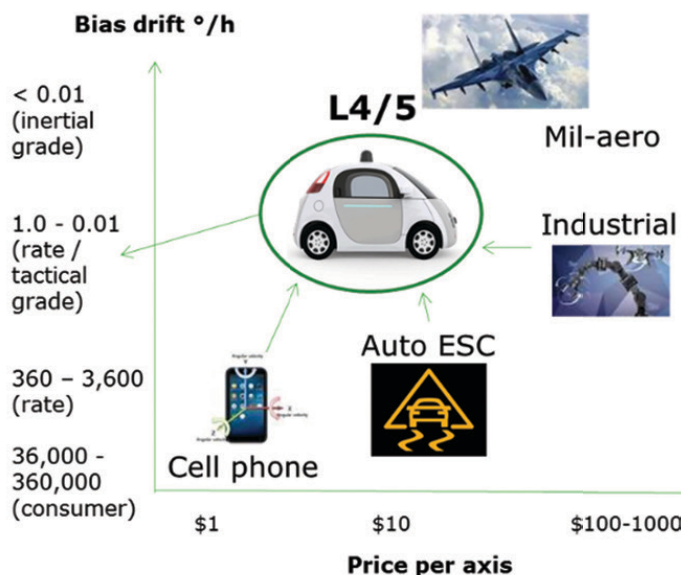


FIGURE 1. Current performance levels of different gyroscopes by application and performance measure in terms of bias drift (IHS Markit).

This poses an interesting potential opportunity for both industrial-performance MEMS-based gyroscope sensor-makers, such as Silicon Sensing Systems, Analog Devices, Murata, Epson Toyocom and TDK InvenSense, and for broader-based sensor component-makers such as Bosch, Panasonic, STMicroelectronics, and TDK (InvenSense and Tronics).

While MEMS can master performance, size and low weight, cost remains the challenge. The fail-operational mode requirement for autonomous driving will accommodate higher prices, at least in the beginning, probably in the \$100+ range at first, even for the relatively low volumes of self-driving cars anticipated by 2030. Nonetheless, automotive volumes are very attractive compared to industrial applications and offer a lucrative future market for dead-reckoning sensors.

Your cockpit will get smarter

Automakers are banking on the idea that people like to control their own physical environment. Interiors already feature force and pressure sensors that provide more personalized seating experiences and advanced two-stage airbags for improved safety. In some vehicles, automakers are using pairs of MEMS microphones for noise reduction and image or MEMS infrared sensors for detection of driver presence. Eventually, we might see gas sensors that monitor in-cabin CO₂ levels, triggering a warning when they detect dangerous levels that could cause drowsiness. These smart sensors would then “tell” the driver to open the window or activate an air-scrubbing system in a more complex solution. While today's CO₂ sensors are still relatively expensive, we may see them designed-in as lower-cost versions come to market.

Future cockpits will need to go beyond such concepts in the lead-up to fully automated driving. Seats could contain sensitive acceleration sensors that measure heart and respiration rates as well as body movement and activity. Other devices could monitor body humidity and temperature.

We need look no further than Murata, a supplier initially targeting hospital beds with a MEMS accelerometer as a replacement for pulse oximeters. That same Murata accelerometer could be placed potentially in a car seat to detect heart rate. It's not the only way to do this: another sensing approach for heart-rate measurement comprises millimeter wave radiation, a method that can even look through objects such as books and magazines.

Augmenting sensor-based body monitoring, automotive designers will use cameras to fuse information such as gaze direction, rate of blinking and eye closure, head tilt, and seat data with data gathered by sensors to provide valuable information on the driver's physical condition, awareness and even mood.

Faurecia's Active Wellness concept—unveiled at the 2016 Paris Motor Show—proves that this technology might be coming sooner than we think. Active Wellness collects and analyzes biological data and stores the driver's behavior and preferences. This prototype provides data to predict driver comfort based on physical condition, time of day, and traveling conditions, as well as car operating modes: L3, L4 or L5. Other features such as event-triggered massage, seat ventilation and even changes in ambient lighting or audio environment are possible (**FIGURE 2**).

Meanwhile, there are other commercial expressions of more advanced HMI as well as plenty of prototypes. Visteon's Horizon cockpit can use voice activation and hand gestures to open and adjust HVAC. Capacitive sensors are already widely used for touch applications, and touchless possibilities range from simple infrared diodes for proximity measurement to sophisticated 3D time-of-flight measurements for gesture control.

Clearly, automotive designers will have a lot more freedom with HMI in the cabin space, providing a level of differentiation that manufacturers think customers will appreciate—and for which they will pay a premium.

Managing sensor proliferation

Researchers are investigating ways to solve the issue of high-functionality vehicles containing myriad sensing inputs, i.e., when we have so many sensing inputs, designers must address wiring complexity and unwanted harness weight. Faurecia, for example, is considering ways to convert wood, aluminum, fabric or plastic into smart surfaces that can be functionalized via touch-sensitive capacitive switches integrated into the surface. These smart surfaces could reduce the explosion of sensing inputs, thereby diminishing wiring complexity. With availability from 2020, Faurecia's solutions are approaching the market soon.

Beyond functionalized switches, flexible electronics and wireless power sources, and even energy harvesting (to mitigate power sources), could provide some answers. Indeed, recent research has shown that graphene-based Hall-effect

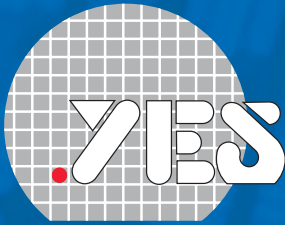


FIGURE 2. Faurecia's "cockpit of the future," announced at CES 2018. (Faurecia).

devices can be embedded in large-area flexible Kapton films, and eventually integrated into panels. OEMs such as Jaguar Land Rover are interested in such approaches to address the downsides of electronics and sensor proliferation, especially in luxury vehicles. While smart surfaces would represent a big change in sensor packaging and a disruption in current semiconductor processes, they remain a long way from commercial introduction.

By 2030 or thereabouts, fully autonomous cars that detect our mood, vital signs and activity level could well be available. Cabins could signal us to open the window if CO2 levels become dangerous. HVAC systems could increase seat ventilation or turn up the air conditioning (or the heat) based on our body temperature. Feeling too hot or too cold in the cabin could become a thing of the past, at least for the driver, whose comfort level is the most important! We could feasibly feel more comfortable in the car than in our office, our home or at the movies. Perhaps our car will become our office, our entertainment center and our home away from home as we take long road trips with the family, without a single passenger uttering, "Are we there yet?"

Editor's Note: This was originally published in the SEMI-MEMS & Sensors Industry Group Blog on www.solid-state.com. ◀



Yield Engineering Systems, Inc.

Revolutionary Productivity for Your Advanced Technologies!

**Advanced Packaging,
MEMS, LED, MicroLED,
Power Devices and Sensors...**

YES-ÉcoClean

Automated Plasma Resist Strip/
Descum System (up to 200mm wafers)

- 2x faster, 1/2 the capital cost, and 1/2 the footprint of comparative products
- Elegantly simple system with low cost ownership
- No defects or damage due to ICP Downstream Plasma
- Eco-friendly “Green” process
- Flexible System from Descum to Strip – 100 to 100,000 Å/min

Yield Engineering Systems, Inc.

Call: **+1 925-373-8353** (worldwide) or

1-888-YES-3637 (US toll free)

www.yieldengineering.com



SMART PUMPS FOR SMART MANUFACTURING



LEVITRONIX® PUMP SYSTEMS

The Only Pump with High Performance
Data Acquisition, Ideal for:

- // Predictive Maintenance
- // Continuous Monitoring of
Equipment Performance
- // Increased System Uptime

LEVITRONIX®

sales@levitronix.com · www.levitronix.com