OAK RIDGE National Laboratory

Preparing for Extreme Heterogeneity in High Performance Computing

Jeffrey S. Vetter
With many contributions from FTG Group and Colleagues

DATE 2019
Special Session on Embedded meets Hyperscale and HPC
27 Mar 2019

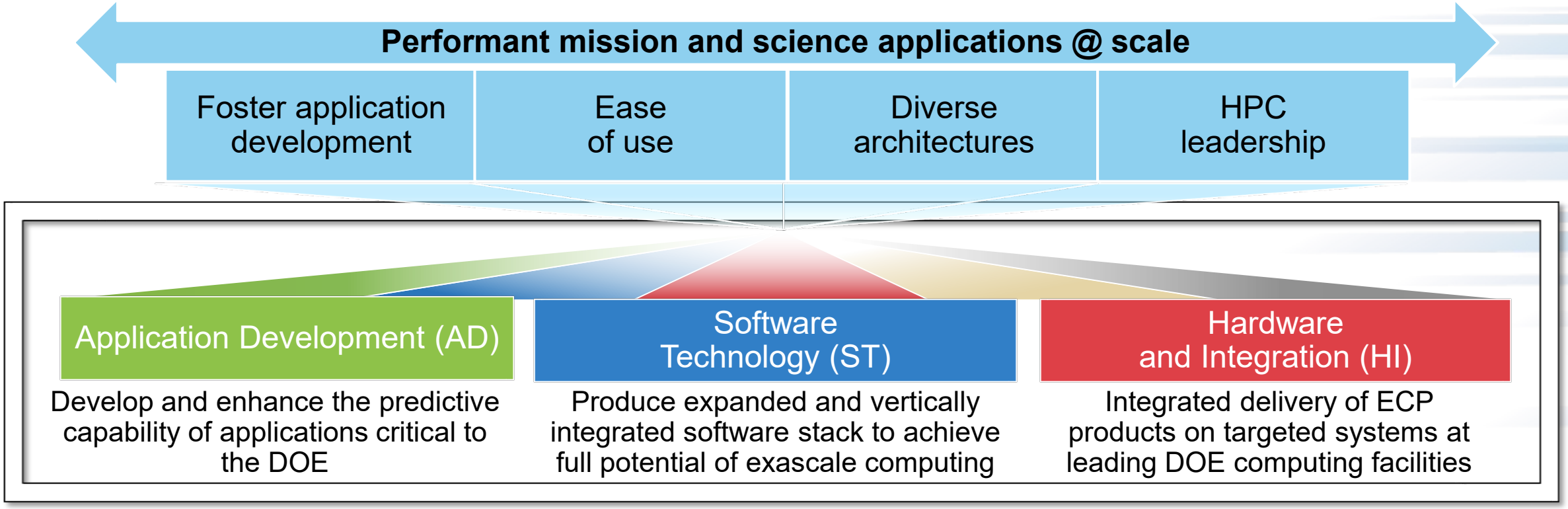
ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Highlights

- Recent trends in extreme-scale HPC paint an uncertain future
  - Contemporary systems provide evidence that power constraints are driving architectures to change rapidly
  - Multiple architectural dimensions are being (dramatically) redesigned: Processors, node design, memory systems, I/O
  - Complexity is our main challenge
- Applications and software systems are all reaching a state of crisis
  - Applications will not be functionally or performance portable across architectures
  - Programming and operating systems need major redesign to address these architectural changes
  - Procurements, acceptance testing, and operations of today's new platforms depend on performance prediction and benchmarking.
- We need portable programming models and performance prediction now more than ever!
- Programming systems must provide performance portability (beyond functional portability)!!
  - Emerging memory hierarchies
    - DRAGON – transparent NVM access from GPUs
    - NVL-C – user management of nonvolatile memory in C
    - Papyrus – parallel aggregate persistent storage
  - Heterogeneous processor (not covered today)
    - OpenACC->FGPAs
    - Clacc – OpenACC support in LLVM
- Performance prediction is critical for design and optimization (not covered today)

The three technical areas in ECP have the necessary components to meet national goals



25 applications ranging from national security, to energy, earth systems, economic security, materials, and data

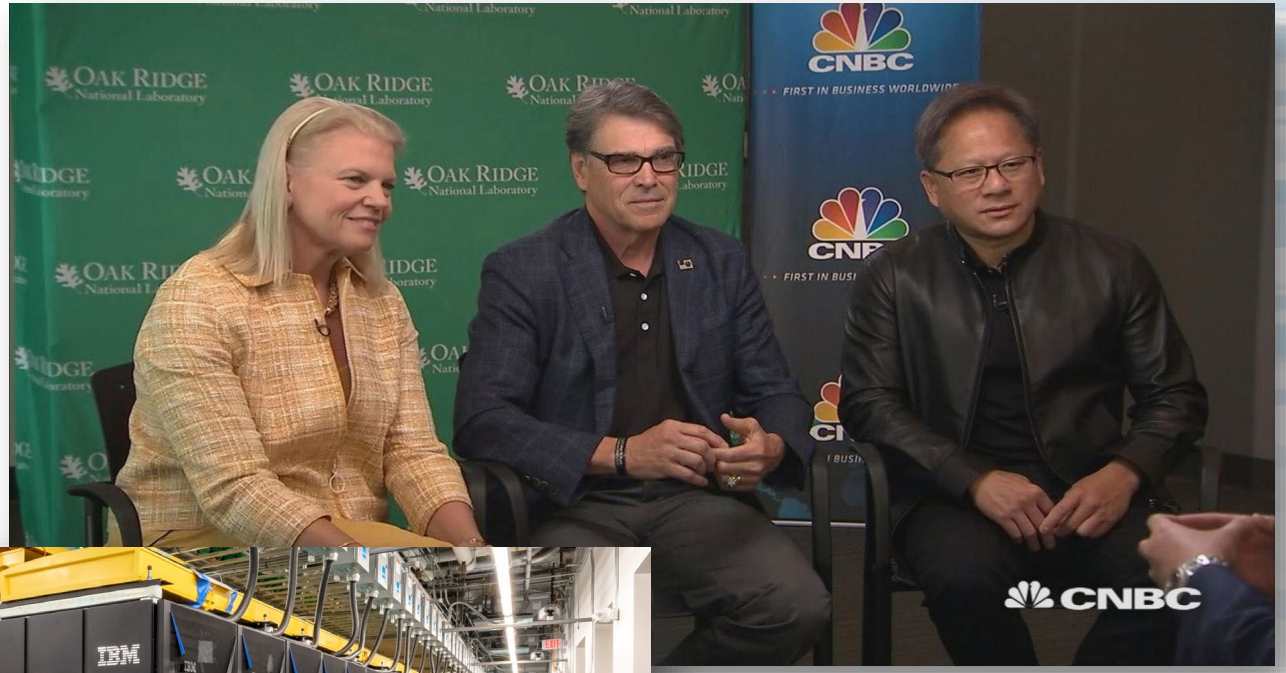
80+ unique software products spanning programming models and run times, math libraries, data and visualization

6 vendors supported by PathForward focused on memory, node, connectivity advancements; deployment to facilities

ORNL 75th Lab Day and Summit Unveiling – 8 June 2018

#1 on Top 500

Application Performance	200 PF
Number of Nodes	4,608
Node performance	42 TF
Memory per Node	512 GB DDR4 + 96 GB HBM2
NV memory per Node	1600 GB
Total System Memory	>10 PB DDR4 + HBM2 + Non-volatile
Processors	2 IBM POWER9™ 9,216 CPUs 6 NVIDIA Volta™ 27,648 GPUs
File System	250 PB, 2.5 TB/s, GPFS™
Power Consumption	13 MW
Interconnect	Mellanox EDR 100G InfiniBand
Operating System	Red Hat Enterprise Linux (RHEL) version 7.4



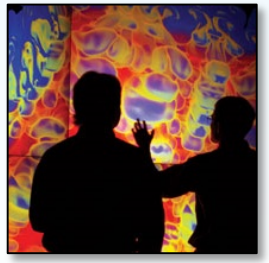
OAK RIDGE National Laboratory | 75 YEARS

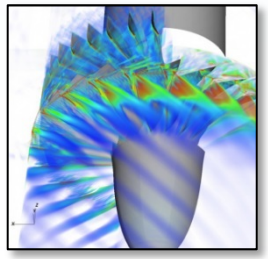
ECP applications target national problems in 6 strategic areas

National security

Stockpile stewardship

Next-generation electromagnetics simulation of hostile environment and virtual flight testing for hypersonic re-entry vehicles





Energy security

Turbine wind plant efficiency

High-efficiency, low-emission combustion engine and gas turbine design

Materials design for extreme environments of nuclear fission and fusion reactors

Design and commercialization of Small Modular Reactors

Subsurface use for carbon capture, petroleum extraction, waste disposal

Scale-up of clean fossil fuel combustion

Biofuel catalyst design

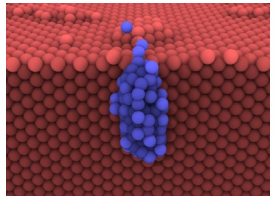
Economic security

Additive manufacturing of qualifiable metal parts

Reliable and efficient planning of the power grid

Seismic hazard risk assessment

Urban planning



Scientific discovery

Find, predict, and control materials and properties

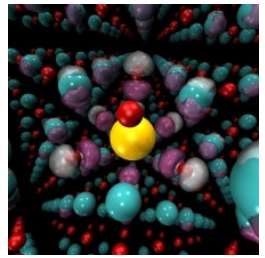
Cosmological probe of the standard model of particle physics

Validate fundamental laws of nature

Demystify origin of chemical elements

Light source-enabled analysis of protein and molecular structure and design

Whole-device model of magnetically confined fusion plasmas



Earth system

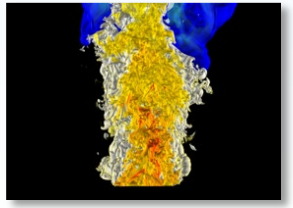
Accurate regional impact assessments in Earth system models

Stress-resistant crop analysis and catalytic conversion of biomass-derived alcohols

Metagenomics for analysis of biogeochemical cycles, climate change, environmental remediation

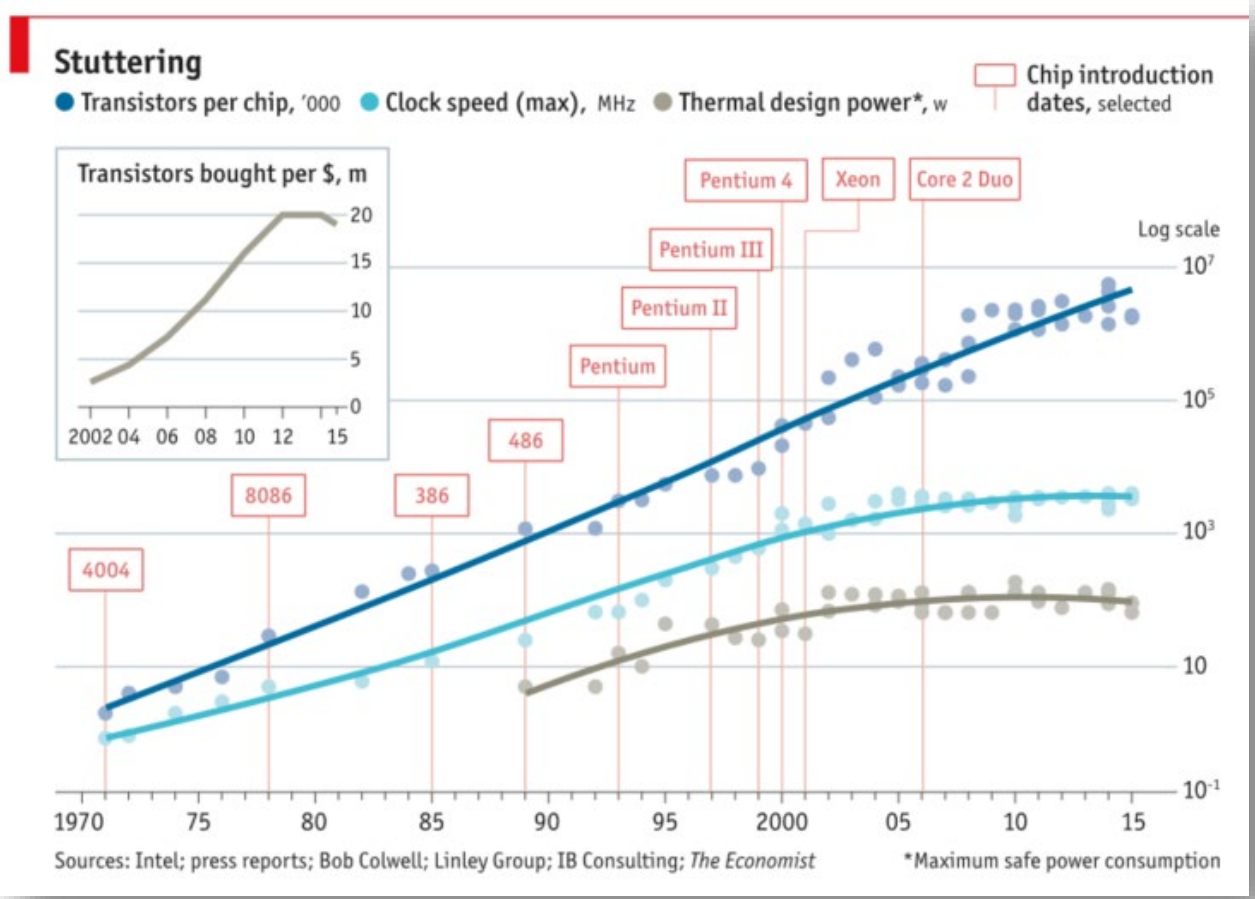
Health care

Accelerate and translate cancer research



Major Trends in Computing

Contemporary devices are approaching fundamental limits



Economist, Mar 2016

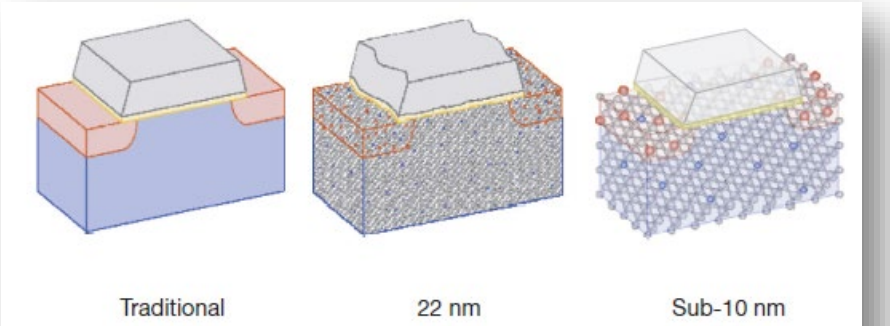


Figure 1 | As a metal oxide–semiconductor field effect transistor (MOSFET) shrinks, the gate dielectric (yellow) thickness approaches several atoms (0.5 nm at the 22-nm technology node). Atomic spacing limits the

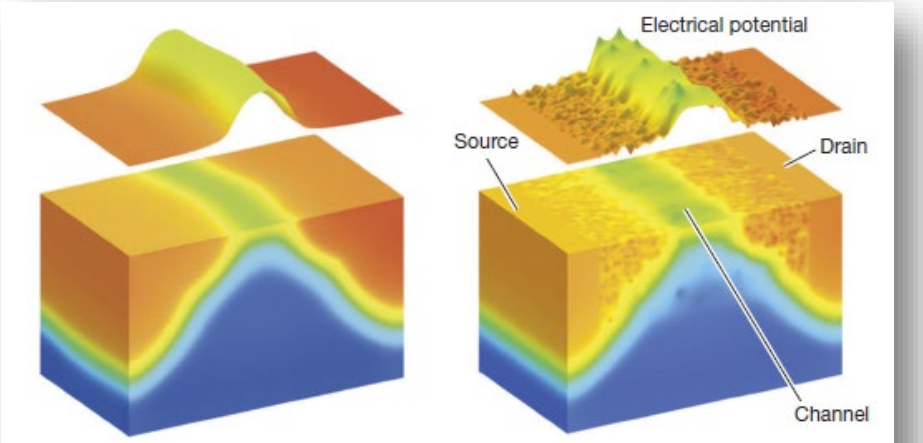


Figure 2 | As a MOSFET transistor shrinks, the shape of its electric field departs from basic rectilinear models, and the level curves become disconnected. Atomic-level manufacturing variations, especially for dopant

Dennard scaling has already ended. Dennard observed that voltage and current should be proportional to the linear dimensions of a transistor: 2x transistor count implies 40% faster and 50% more efficient.

R.H. Dennard, F.H. Gaensslen, V.L. Rideout, E. Bassous, and A.R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits, 9(5):256-68, 1974,*

I.L. Markov, "Limits on fundamental limits to computation," *Nature, 512(7513):147-54, 2014, doi:10.1038/nature13570.*

Business climate reflects this uncertainty, cost, complexity, consolidation

designlines WIRELESS & NETWORKING

Blog

IC Merger Mania Hits Fever Pitch

Dylan McGrath, Contributing Editor
12/2/2015 10:13 AM EST

NO RATINGS
LOGIN TO RATE

1 comments post a c

Like 10 Twee

With the announcem
PMC-Sierra, the total
acquisitions announ

The wave of consolid

Intel to acquire Altera for $54 a share

Monday, 1 Jun 2015 | 8:33

REUTERS

Avago Agrees to Buy Broadcom for $37 Billion

By MICHAEL J. de la MERCED and CHAD BRAY MAY 28, 2015

SANDISK COMPLETES ACQUISITION OF FUSION IO

JUL 23, 2014

ACQUISITION TO BOOST SANDISK'S ENTERPRISE GROWTH

MILPITAS, Calif., July 23, 2014 - SanDisk Corporation (NASDAQ: SNDK), a global leader in flash storage solutions, today announce
hardware

"I am deli
the Fusio
and chief
the indus

to-market talent of
ehrotra, president
se flash solutions in

Western Digital Now A Storage Powerhouse With SanDisk Acquisition

designlines AUTOMOTIVE

News & Analysis

Foundries' Sales Show Hard Times Continuing

SEMICONDUCTOR ENGINEERING

Home > Manufacturing, Design & Test > Uncertainty Grows For 5nm, 3nm

MANUFACTURING, DESIGN & TEST

Uncertainty Grows For 5nm, 3nm

797 74

Nanosheets and nanowire FETs under development, but costs are skyrocketing. New packaging options could provide an alternative.

DECEMBER 19TH, 2016 - BY: MARK LAPEDUS

As several chipmakers ramp up their
processes, with 7nm just around the
for 5nm and beyond. In fact, some ar
speed ahead in the arena.

TSMC recently announced plans to b
at a cost of $15.7 billion. The propos

er Clarke
2016 09:33 PM EDT
mments

Like 6 Tweet

DON--Taiwanese found

Tech giant ARM Holdings sold to Japanese firm for £24bn

Britain's largest tech firm,
deal including UK jobs gu

Qualcomm to Acquire NXP Semiconductors for $38.5 Billion

By CHAD BRAY and QUENTIN HARDY OCT. 27, 2016

SoftBank to sell 25% of Arm to Saudi-backed fund

Son puts stake worth $8bn in UK's largest tech company into $100bn Vision Fund

Broadcom acquires Brocade in $5.9 billion deal

Posted 1 hour ago by Ron Miller (@ron_miller)

Next Story

#BUSINESS NEWS NOVEMBER 19, 2017 / 7:57 PM / UPDATED 21 MINUTES AGO

Marvell Technology to buy rival chipmaker Cavium for $6 billion

Toshiba to sell 'minority sta Western Digital

In April-June 2016, Toshiba had a 20.4% share in global NAN

usiness Bureau

nglomerate Toshiba Corporation is in d
lash memory division to the US maker
ing to a source

designlines SoC

News & Analysis

TSMC Grows Share of Foundry Business

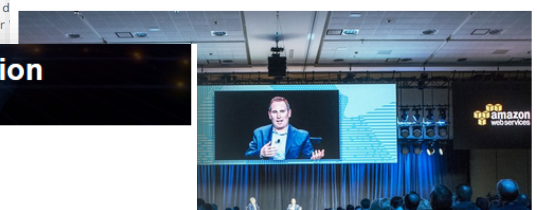
ell Technology Group Ltd (MRV

EXCLUSIVE

Amazon Is Becoming an AI Chip Maker, Speeding Alexa Responses

TOM SIMONITE BUSINESS 11.27.18 08:12 PM

NEW AT AMAZON: ITS OWN CHIPS FOR CLOUD COMPUTING



Repercussions of Samsung's b

Alan Patterson
10/13/2016 09:38 AM EDT
Post a comment

Tweet Share 20 G+

TAIPEI — Taiwan Semiconductor Manufac

eetasia.com

GlobalFoundries Forfeit 7nm Manufacturing - EE Times Asia

6-7 minutes

SAN JOSE, Calif. – The race to drive semiconductor technology to

:00 AM PST

ping a chip desi
l other hardwar
iliar with Amaz
spond more qu
handled on the
e latest major te
hips, in hopes

In Intel's Arduous Journey to 10 nm, Moore's Law Comes Up Short

Dairsie Latimer, Technical Advisor, Red Oak Consulting | August 30, 2018 11:53 CEST

E-mail Tweet Like

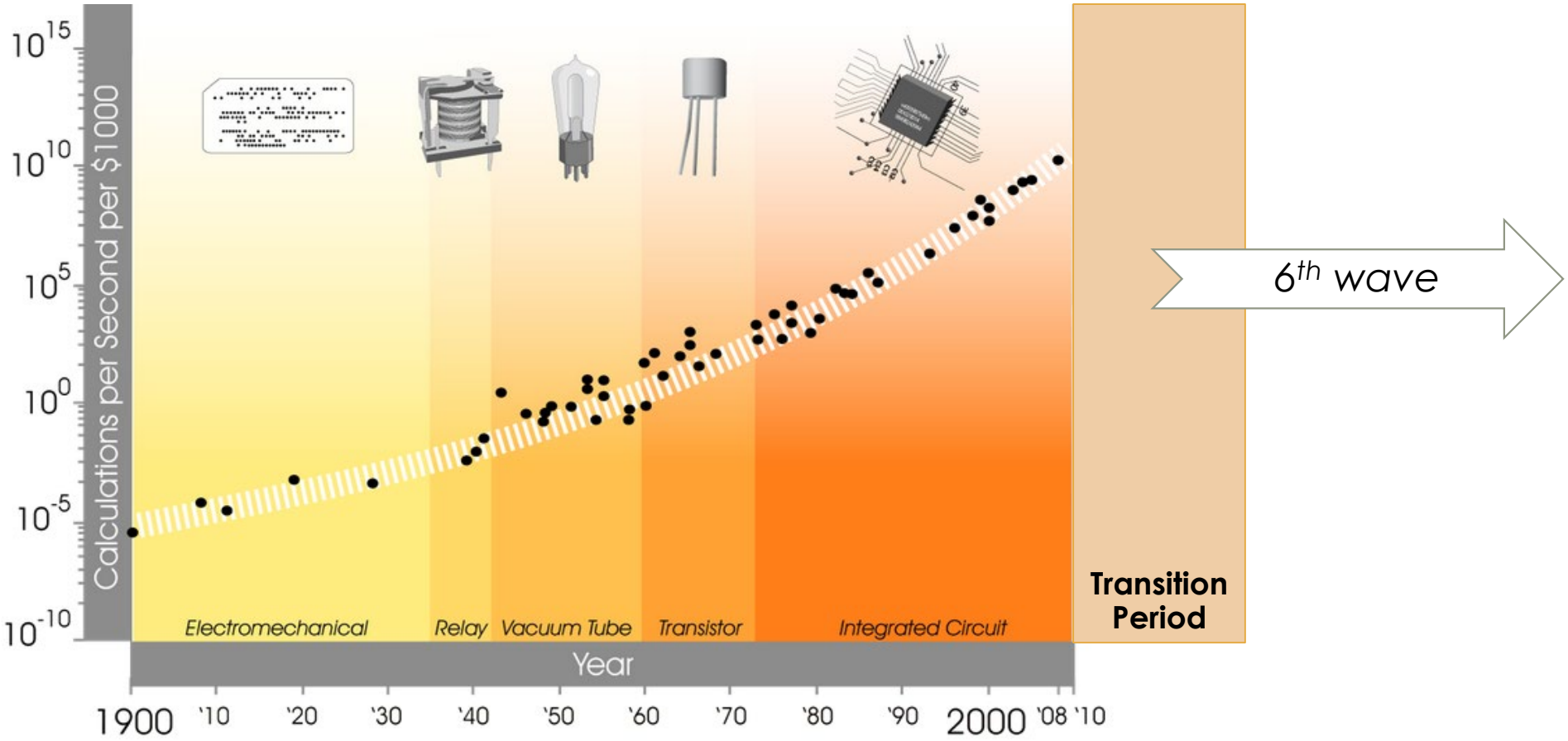
With a share price riding high and dominan
facing a range of significant problems. So v
itself so spectacularly on its back foot?

Nvidia Wins Mellanox Stakes for $6.9 Billion

By Doug Black

March 11, 2019

Sixth Wave of Computing



http://www.kurzweilai.net/exponential-growth-of-computing

Transition Period Predictions

Optimize Software and Expose New Hierarchical Parallelism

- Redesign software to boost performance on upcoming architectures
- Exploit new levels of parallelism and efficient data movement

Architectural Specialization and Integration

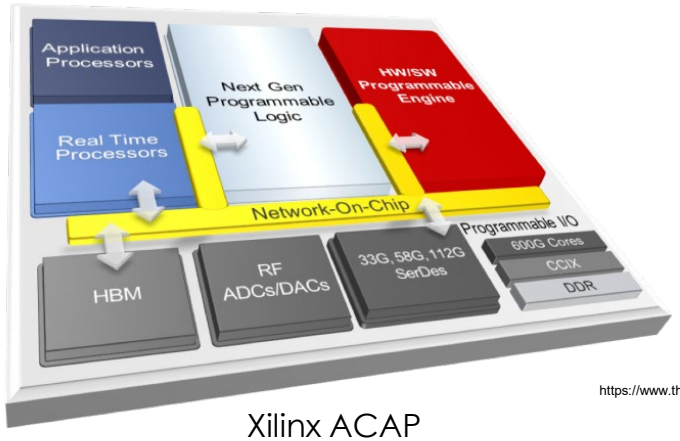
- Use CMOS more efficiently for our workloads
- Integrate components to boost performance and eliminate inefficiencies

Emerging Technologies

- Investigate new computational paradigms
  - Quantum
  - Neuromorphic
  - Advanced Digital
  - Emerging Memory Devices

Architectural specialization is quickening

- Vendors, lacking Moore's Law, will need to continue to differentiate products (to stay in business)
- Grant that advantage of better CMOS process stalls
- Use the same transistors differently to enhance performance
- Architectural design will become extremely important, critical
  - Dark Silicon
  - Address new parameters for benefits/curse of Moore's Law



Xilinx ACAP

https://www.thebroadcastbridge.com/content/entry/1094/altera-announces-arria-10-2666mbps-ddr4-memory-fpga-interf

Intel's Nervana AI platform takes aim at Nvidia's GPU techology

Firm claims Xeon-based chips will deliver a '100-fold increase' in deep learning performance

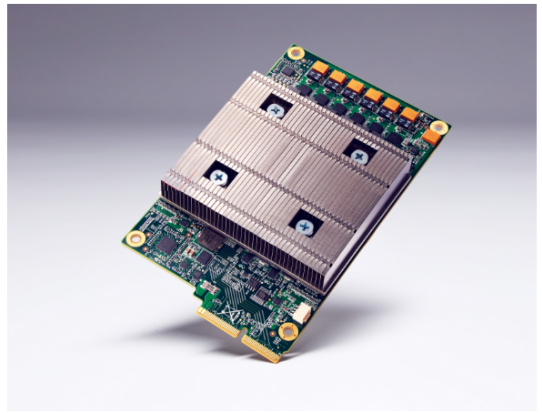
CHIPMAKER Intel has set out its plans for artificial intelligence (AI) and claimed that it will reduce the time to train a deep learning model by up to 100 times within the next three years.

At the forefront of the firm's AI ambitions is the Intel Nervana platform, which was announced on Thursday **following Intel's acquisition of deep learning startup Nervana Systems earlier this year**.

http://www.theinquirer.net/inquirer/news/2477796/intels-nervana-ai-platform-takes-aim-at-nvidias-gpu-techology

CADE METZ BUSINESS 05.18.16 3:57 PM

GOOGLE BUILT ITS VERY OWN CHIPS TO POWER ITS AI BOTS

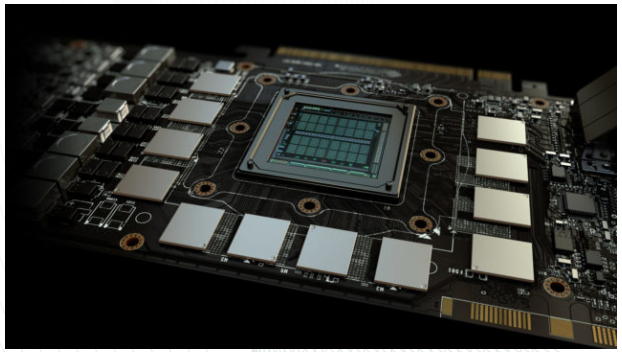


GOOGLE

GOOGLE HAS DESIGNED its own computer chip for driving deep neural networks, an AI technology that is reinventing the way Internet services operate.

This morning at Google I/O, the centerpiece of the company's year, CEO Sundar Pichai said that Google has designed an ASIC, or application-specific integrated circuit, that's specific to deep neural nets. These are networks of

http://www.wired.com/2016/05/google-tpu-custom-chips/



https://fossbytes.com/nvidia-volta-gddr6-2018/

D.E. Shaw, M.M. Deneroff, R.O. Dror *et al.*, *"Anton, a special-purpose machine for molecular dynamics simulation," Communications of the ACM, 51(7):91-7, 2008.*

Turing Award Lecture on June 4: A New Golden Age for Computer Architecture

- Domain-specific HW/SW Co-Design
- Enhanced Security
- Open Instruction Sets
- Agile Chip Development

A New Golden Age for Computer Architecture: Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development

John L. Hennessy and David A. Patterson

In the 1980s, Mead and Conway[1] democratized chip design and high-level language programming surpassed assembly language programming, which made instruction set advances viable. Innovations like RISC, superscalar, multilevel caches, and speculation plus compiler advances (especially in register allocation) ushered in a Golden Age of computer architecture, when performance increased annually by 60%. In the later 1990s and 2000s, architectural innovation decreased, so performance came primarily from higher clock rates and larger caches. The ending of Dennard Scaling and Moore's Law also slowed this path; single core performance improved only 3% last year! In addition to poor performance gains of modern microprocessors, Spectre recently demonstrated timing attacks that leak information at high rates[2].

We're on the cusp of another Golden Age that will significantly improve cost, performance, energy, and security. These architecture challenges are even harder given that we've lost the exponentially increasing resources provided by Dennard scaling and Moore's law. We've identified areas that are critical to this new age:

1. **Hardware/Software Co-Design for High-Level and Domain-Specific Languages**
   Advanced programming languages like Python and domain-specific languages like TensorFlow have dramatically improved programmer productivity by increasing software reuse and by raising the level of abstraction. Whereas compiler-architecture co-design delivered gains of about three in the 1980s for C compilers and RISC architectures, new advances could create compilers and domain-specific architectures[3] (DSAs) that deliver tenfold or more jumps[4] in this new Golden Age.

2. **Enhancing Security**
   We've made tremendous gains in information technology (IT) in the past 40 years, but if security is a war, we're losing it. Thus far, architects have been asked for little beyond page-level protection and supporting virtual machines. The very definition of computer architecture ignores timing, yet Spectre shows that attacks that can determine timing of operations can leak supposedly protected data. It's time for architects to redefine computer architecture and treat security as a first class citizen to protect data from timing attacks, or at worst reduce information leaks to a trickle.

3. **Free and Open Architectures and Open-Source Implementations**
   Progress on these issues likely will require changes to the instruction set architecture (ISA), which is problematic for proprietary ISAs. For tall challenges like these, we want all the best minds to work on them, not only the engineers who work for the ISA owners. Thus, a free and open ISA such as RISC-V can be a boon to researchers[5] because:

   - Many people in many organizations can innovate simultaneously using RISC-V.
   - The ISA is designed for modularity and extensions.
   - It comes with a complete software stack, including compilers, operating systems, and debuggers, which are open source and thus also modifiable.
   - This modern ISA is designed to work for any application, from cloud-level servers down to mobile and IoT devices.
   - RISC-V is driven by a 100-member foundation[6] that ensures its long-term stability and evolution.

   Unlike the past, open ISAs are viable because many engineers for a wide range of products are designing SOCs by incorporating IP and because ARM has demonstrated that IP works for ISAs.

   An open architecture also enables open-source processor designs for both FPGAs and real chips, so architects can innovate by modifying an existing RISC-V design and its software stack. While FPGAs run at perhaps only 100 MHz, that is fast enough to run trillions of instructions or to be deployed on the Internet to test a security feature against real attacks. Given the plasticity of FPGAs, the RISC-V ecosystem enables experimental investigations of novel features that can be deployed, evaluated, and iterated in days rather than in years. That vision requires more IP than CPUs, such as GPUs, neural network accelerators, DRAM controllers, and PCIe controllers[7]. The stability of process nodes due to the ending of Moore's Law make this goal easier than in the past. This necessity opens a path for architects to have impact by contributing open-source components much as their software colleagues do for databases and operating systems.

4. **Agile Chip Development**
   As the focus of innovation in architecture shifts from the general-purpose CPU to domain-specific and heterogeneous processors, we will need to achieve major breakthroughs in design time and cost (as happened for VLSI in the 1980s). Small teams should be able to design chips, tailored for a specific domain or application. This will require that hardware design become much more efficient, and more like modern software design.

   Unlike the "waterfall" development process of giant chips by large companies, Agile development process[8] allows small groups to iterate designs of working but incomplete prototypes for small chips. Fortuitously, the same programming language advances that improved reuse of software have been incorporated in recent hardware design languages, which makes hardware design and reuse easier. While one can stop at layout for a research paper, building real chips is inspiring for everyone in a project, and is the only way to verify important characteristics like timing and energy consumption. The good news is that today TMSC will deliver 100 small test chips in the latest technology for only $30,000[9]. Thus, virtually all projects can afford real chips as final validation of innovation as well as to enjoy the satisfaction of seeing your ideas work in silicon.

   We believe the deceleration of performance gains for standard microprocessors, the opportunities in high-level, domain-specific languages and security, the freeing of architects from the chains of proprietary ISAs, and (ironically) the ending of Dennard scaling and Moore's law will lead to another Golden Age for architecture. Aided by an open-source ecosystem, agily developed prototypes will demonstrate advances and thereby accelerate commercial adoption. We envision the same rapid improvement as in the last Golden Age, but this time in cost, energy, and security as well in performance.

Transition Period will be Disruptive

- New devices and architectures may not be hidden in traditional levels of abstraction
  - A new type of CNT transistor may be completely hidden from higher levels
  - A new paradigm like quantum may require new architectures, programming models, and algorithmic approaches
- Solutions need a co-design framework to evaluate and mature specific technologies

Layer	Switch, 3D	NVM	Approximate	Neuro	Quantum
Application	1	1	2	2	3
Algorithm	1	1	2	3	3
Language	1	2	2	3	3
API	1	2	2	3	3
Arch	1	2	2	3	3
ISA	1	2	2	3	3
Microarch	2	3	2	3	3
FU	2	3	2	3	3
Logic	3	3	2	3	3
Device	3	3	2	3	3

Adapted from IEEE Rebooting Computing Chart

HPC Architectures Reflect these Trends

Department of Energy (DOE) Roadmap to Exascale Systems

An impressive, productive lineup of *accelerated node* systems supporting DOE's mission

Pre-Exascale Systems [Aggregate Linpack (Rmax) = 323 PF!]

First U.S. Exascale Systems

2012	2016	2018	2020	2021-2023

ORNL
Cray/AMD/NVIDIA

ORNL
IBM/NVIDIA

Heterogeneous Cores

ORNL
TBD

ANL
IBM BG/Q

ANL
Cray/Intel KNL

Deep Memory incl NVM

ANL
Intel/Cray

LBNL
Cray/Intel Xeon/KNL

LBNL
Cray/AMD/NVIDIA

Plateauing I/O Performance

LLNL
IBM BG/Q

LANL/SNL
Cray/Intel Xeon/KNL

LLNL
IBM/NVIDIA

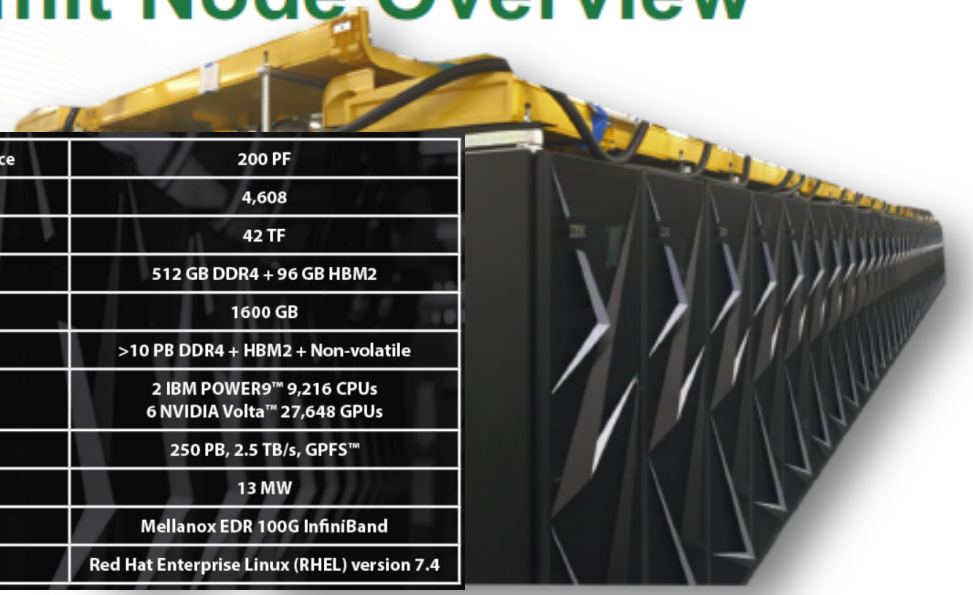
LANL/SNL
TBD

LLNL
TBD

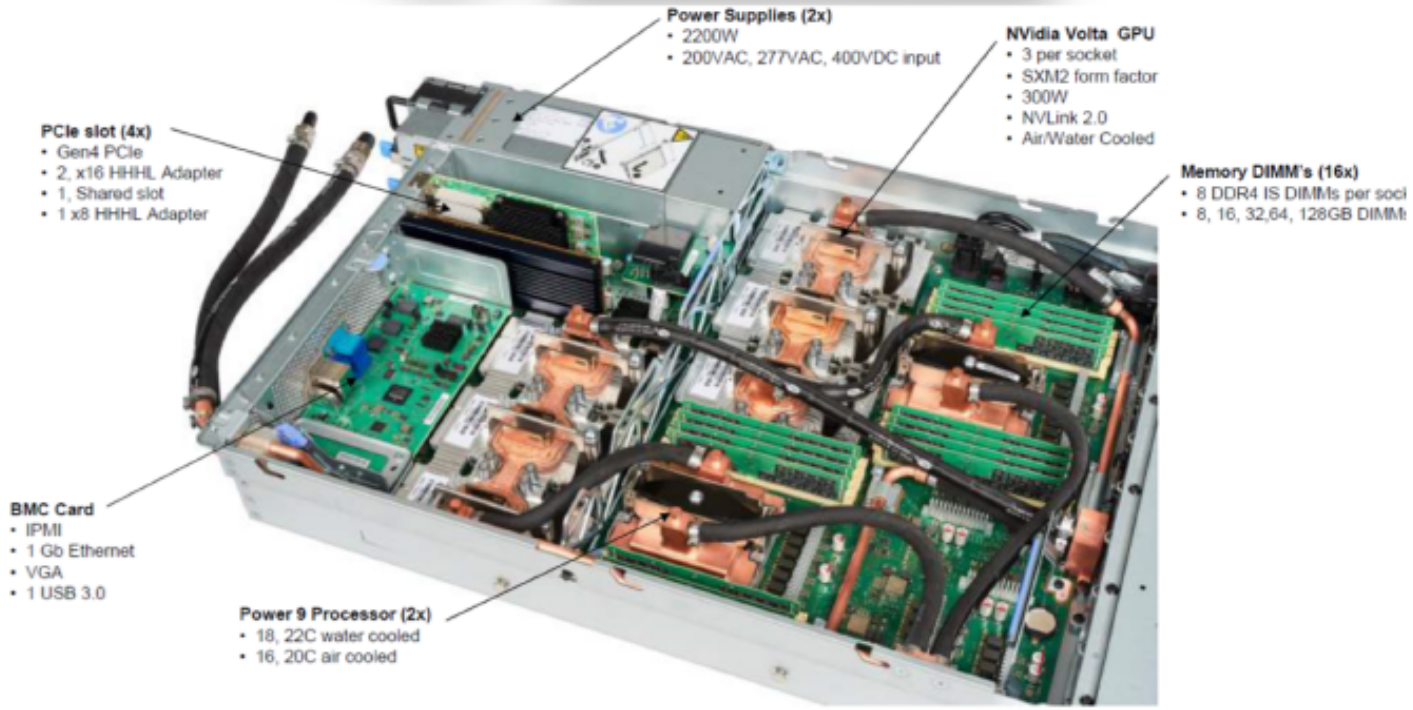
OAK RIDGE National Laboratory

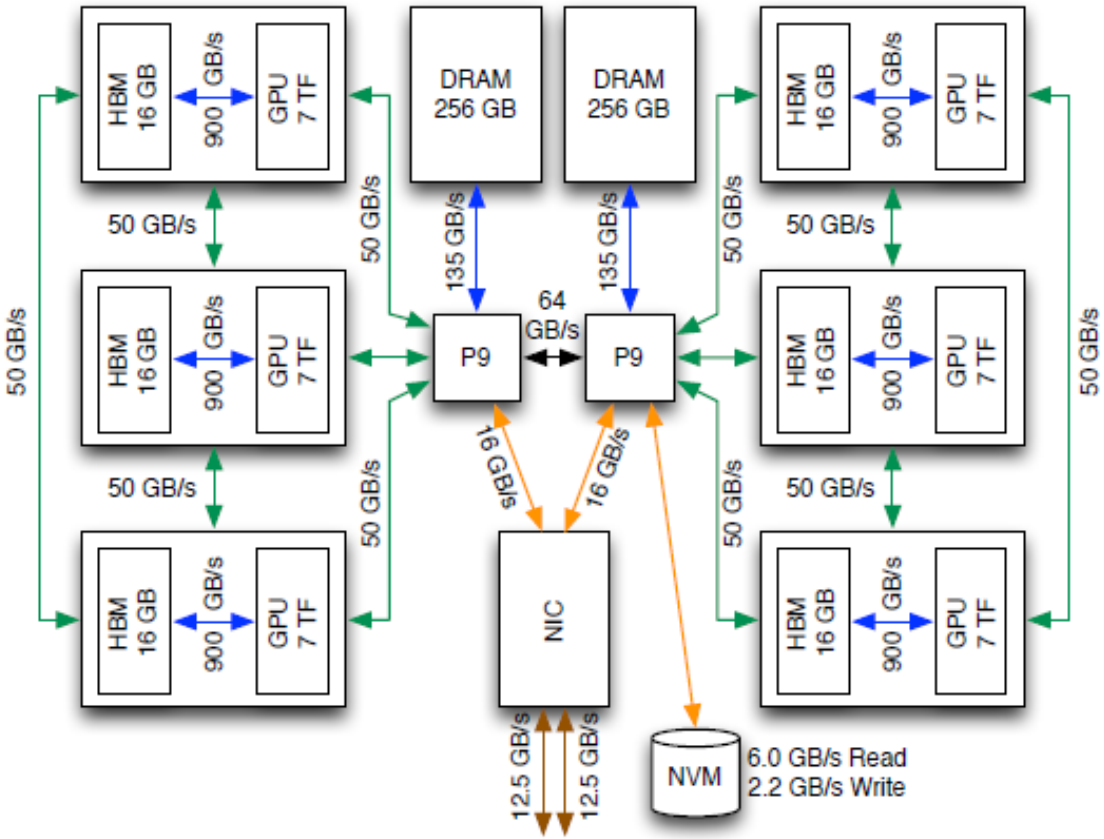
Jan 2018

Summit Node Overview



Application Performance	200 PF
Number of Nodes	4,608
Node performance	42 TF
Memory per Node	512 GB DDR4 + 96 GB HBM2
NV memory per Node	1600 GB
Total System Memory	>10 PB DDR4 + HBM2 + Non-volatile
Processors	2 IBM POWER9™ 9,216 CPUs 6 NVIDIA Volta™ 27,648 GPUs
File System	250 PB, 2.5 TB/s, GPFS™
Power Consumption	13 MW
Interconnect	Mellanox EDR 100G InfiniBand
Operating System	Red Hat Enterprise Linux (RHEL) version 7.4





TF	42 TF (6x7 TF)
HBM	96 GB (6x16 GB)
DRAM	512 GB (2x16x16 GB)
NET	25 GB/s (2x12.5 GB/
MMsg/s	83

HBM/DRAM Bus (aggregate B/W)
NVLINK
X-Bus (SMP)

HBM & DRAM speeds are aggregate (Read+Write).
All other speeds (X-Bus, NVLink, PCIe, IB) are bi-directional.

During this Sixth Wave transition, Complexity is our major challenge!

Design: How do we design future systems so that they are better than current systems on mission applications?

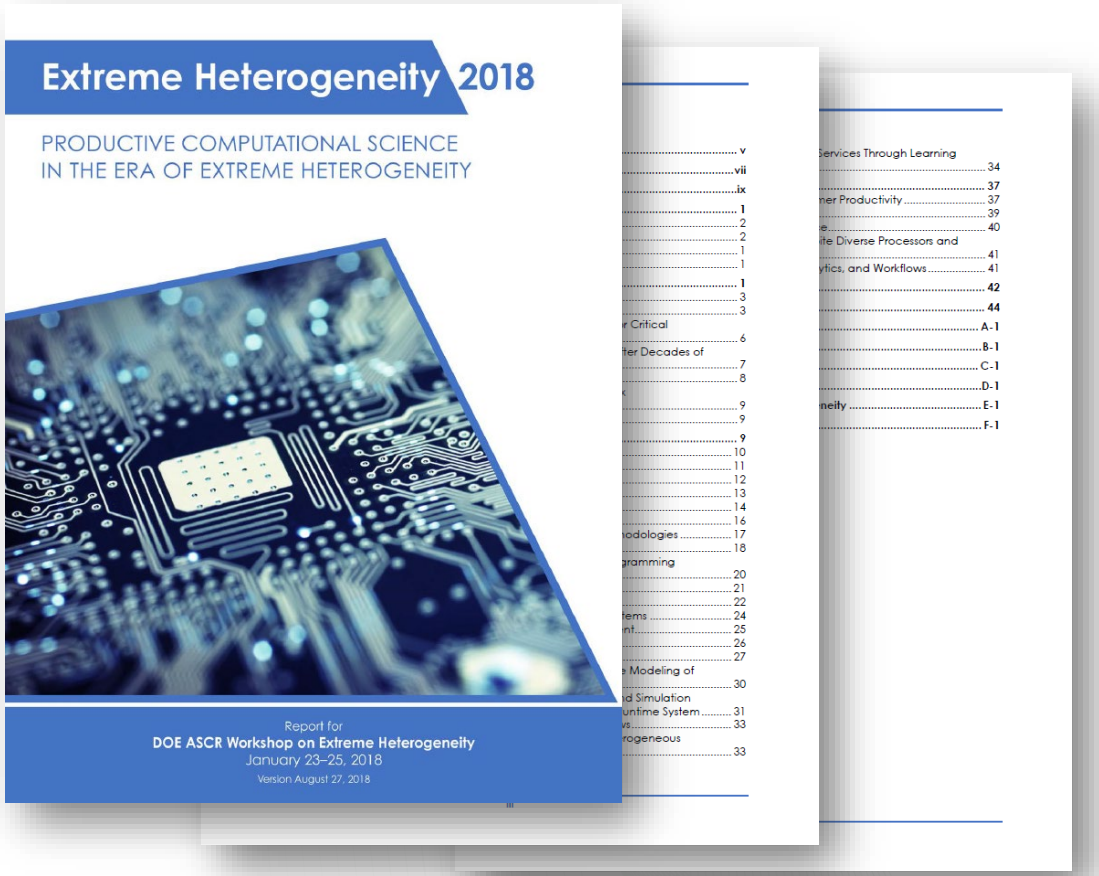
- Entirely possible that the new system will be slower than the old system!
- Expect 'disaster' procurements

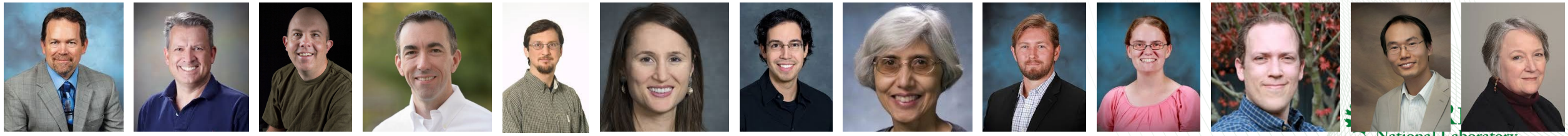
Programmability: How do we design applications with some level of performance portability?

- Software lasts much longer than transient hardware platforms
- Adapt or die

Final Report on Workshop on Extreme Heterogeneity

1. Maintaining and improving programmer productivity
   - Flexible, expressive, programming models and languages
   - Intelligent, domain-aware compilers and tools
   - Composition of disparate software components
- Managing resources intelligently
   - Automated methods using introspection and machine learning
   - Optimize for performance, energy efficiency, and availability
- Modeling & predicting performance
   - Evaluate impact of potential system designs and application mappings
   - Model-automated optimization of applications
- Enabling reproducible science despite non-determinism & asynchrony
   - Methods for validation on non-deterministic architectures
   - Detection and mitigation of pervasive faults and errors
- Facilitating Data Management, Analytics, and Workflows
   - Mapping of science workflows to heterogeneous hardware and software services
   - Adapting workflows and services to meet facility-level objectives through learning approaches





https://orau.gov/exheterogeneity2018/

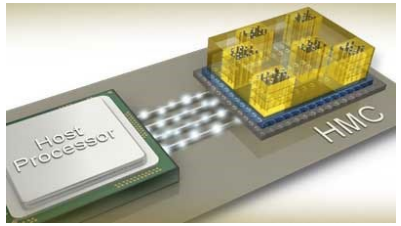
https://doi.org/10.2172/1473756

OAK RIDGE National Laboratory

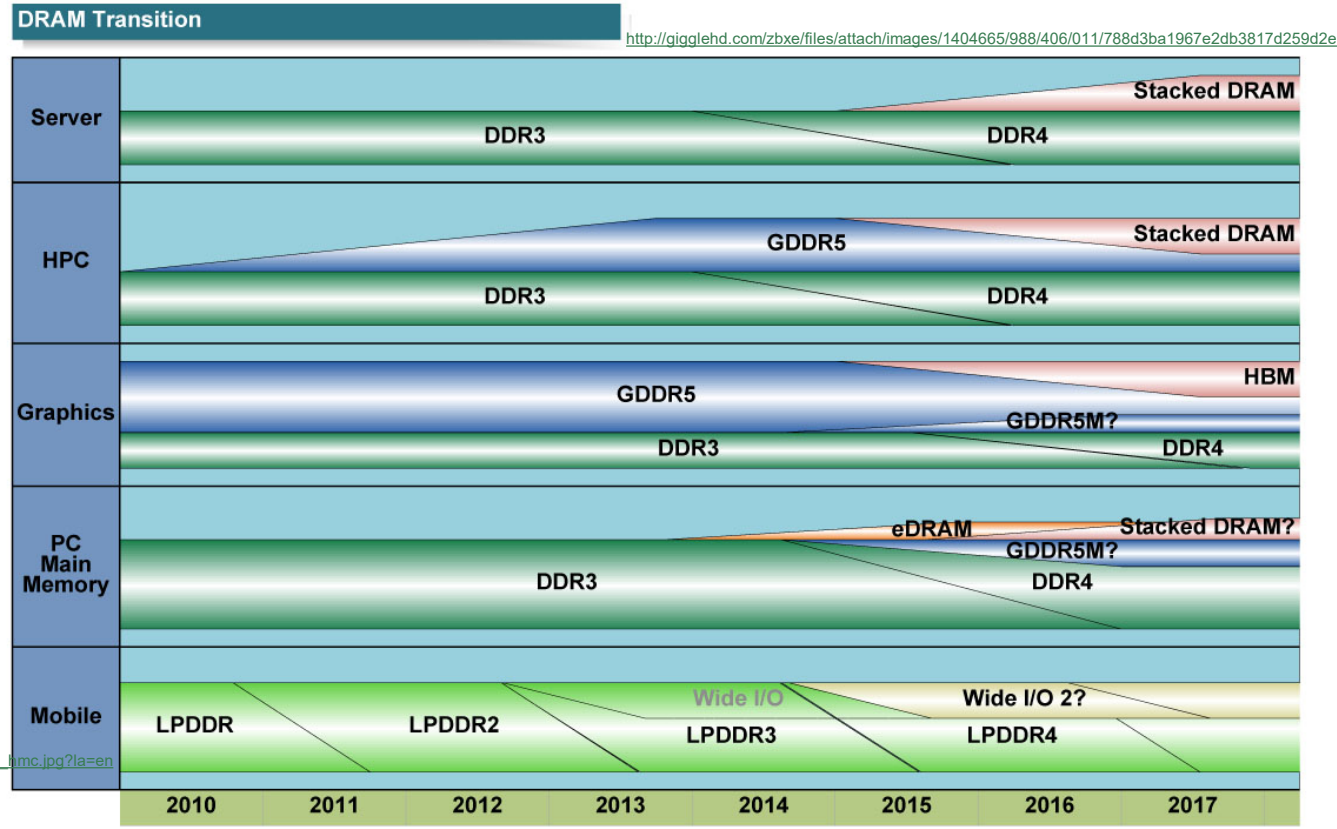
Emerging Memory Systems

Memory Systems Started Diversifying Several Years Ago

- Architectures
  - HMC, HBM/2/3, LPDDR4, GDDR5X, WIDEIO2 etc
  - 2.5D, 3D Stacking
- Configurations
  - Unified memory
  - Scratchpads
  - Write through, write back, etc
  - Consistency and coherence protocols
  - Virtual v. Physical, paging strategies
- New devices
  - ReRAM, PCRAM, STT-MRAM, 3D-Xpoint



https://www.micron.com/~/media/track-2-images/content-images/content_image_hmc.jpg?la=en



http://gigglehd.com/zbxe/files/attach/images/1404665/988/406/011/788d3ba1967e2db3817d259d2e

	SRAM	DRAM	eDRAM	2D NAND Flash	3D NAND Flash	PCRAM	STTRAM	2D ReRAM	3D ReRAM
Data Retention	N	N	N	Y	Y	Y	Y	Y	Y
Cell Size (F^2)	50-200	4-6	19-26	2-5	<1	4-10	8-40	4	<1
Minimum F demonstrated (nm)	14	25	22	16	64	20	28	27	24
Read Time (ns)	< 1	30	5	10^4	10^4	10-50	3-10	10-50	10-50
Write Time (ns)	< 1	50	5	10^5	10^5	100-300	3-10	10-50	10-50
Number of Rewrites	10^{16}	10^{16}	10^{16}	10^4-10^5	10^4-10^5	10^8-10^{10}	10^{15}	10^8-10^{12}	10^8-10^{12}
Read Power	Low	Low	Low	High	High	Low	Medium	Medium	Medium
Write Power	Low	Low	Low	High	High	High	Medium	Medium	Medium
Power (other than R/W)	Leakage	Refresh	Refresh	None	None	None	None	Sneak	Sneak
Maturity									

J.S. Vetter and S. Mittal, "Opportunities for Nonvolatile Memory Systems in Extreme-Scale High Performance Computing," CiSE, *17(2):73-82, 2015.*

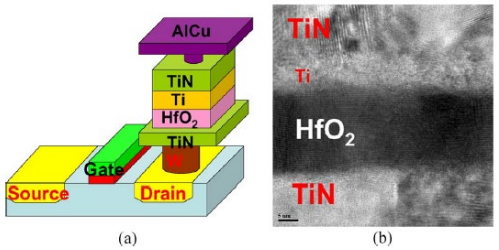


Fig. 4. *(a) A typical 1T1R structure of RRAM with HfO_x; (b) HR-TEM image of the TiN/Ti/HfO_x/TiN stacked layer; the thickness of the HfO_2 is 20 nm.*

H.S.P. Wong, H.Y. Lee, S. Yu *et al.*, *"Metal-oxide RRAM," Proceedings of the IEEE, 100(6):1951-70, 2012.*

Complexity in the Expanding and Diversifying Memory Hierarchy

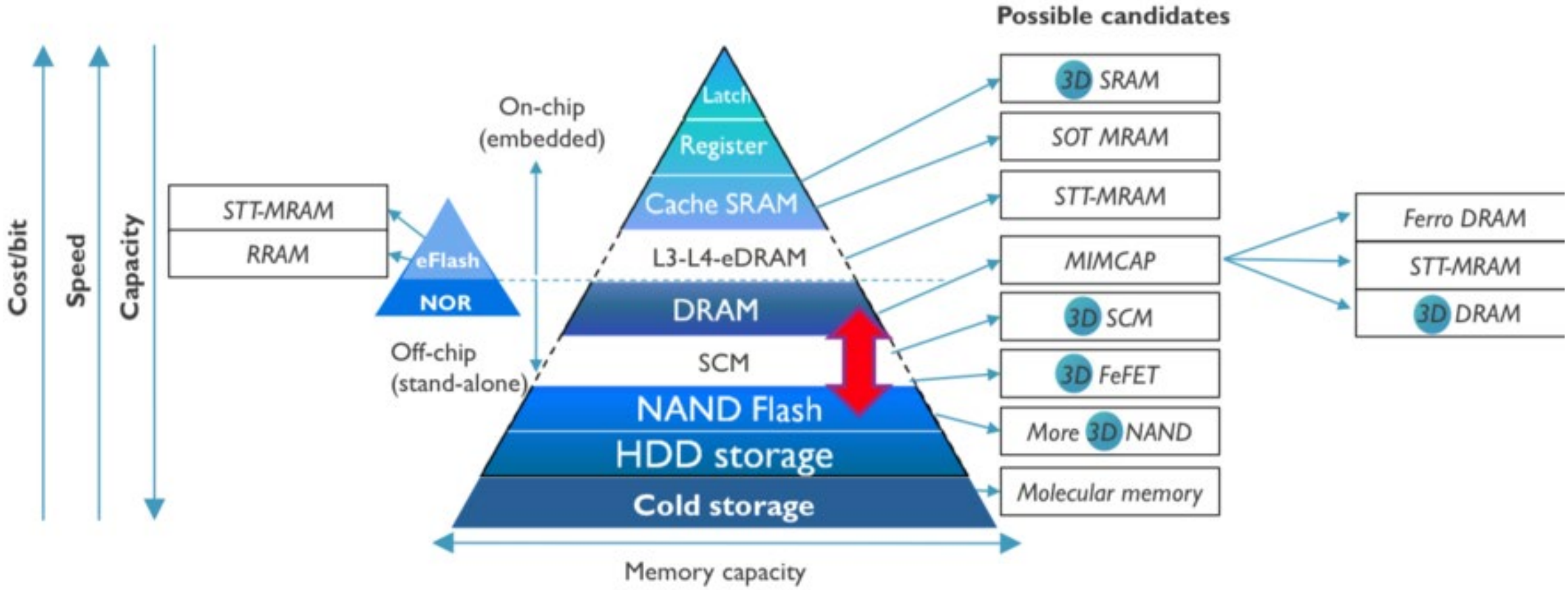


Image Source: IMEC

Many Memory Architecture Options under Consideration…

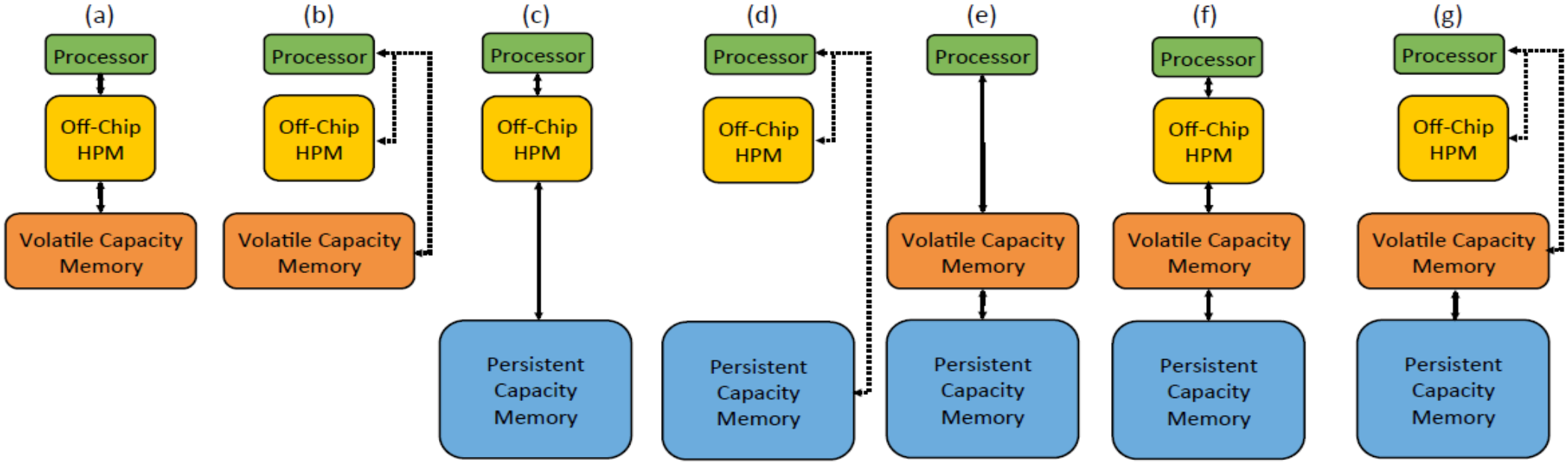


TABLE I: Comparison of four tiers of recent memory technologies [9], [11], [17], [18], [22]–[25], [28], [30], [35], [39], [40], [47]–[49].

	Volatile	Density (GB)	BW (GB/s)	Est. Cost	Speed	Latency
HMC2.0	✓	4-8	320	3x	30 Gbps	∼100s ns
HBM2	✓	2-8	256	2x	2 Gbps	∼100s ns
GDDR6	✓	8-16	72	2x	18 Gbps	∼100s ns
WIO2	✓	8-32	68	2x	1,066 MT/s	∼100s ns
DDR4	✓	2-16	25.6	1x	3,200 MT/s	20-50 ns
STT-MRAM	✗	0.5	-	1x	1,600 MT/s	10-50 ns
PCM	✗	1	3.5	1x	3M IOPS	50-100 ns
3D-Xpoint	✗	750	2.4	0.5x	550K IOPS	10 μs
Z-NAND	✗	800	3.2	0.5x	750K IOPS	12-20 μs
NAND Flash	✗	>1,000	<3	0.1x	50K IOPS	25-125 μs

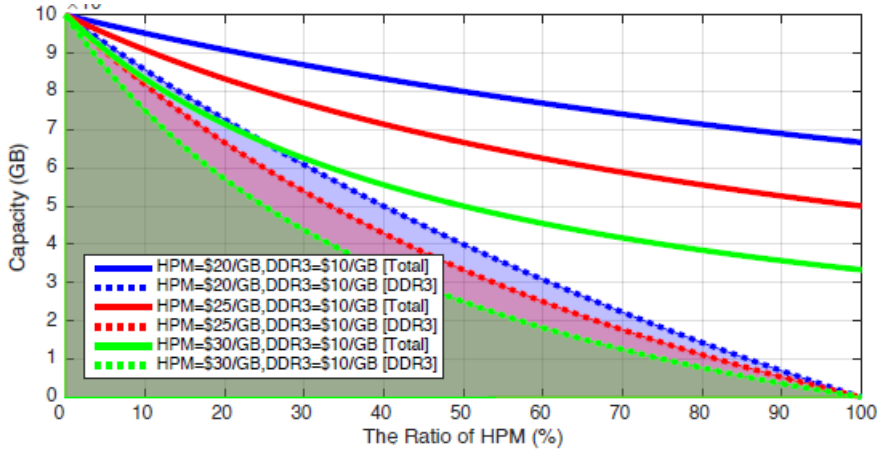
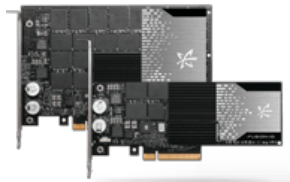


Fig. 1: Possible configurations of a memory system using DDR3 and HPM of different costs under a fixed budget.

I.B. Peng et al, "Siena: Exploring the Design Space of Heterogeneous Memory Systems," in *SC18, 2018*

NVRAM Technology Continues to Improve – Driven by Broad Market Forces



designlines MEMORY

Blog

First Look at Samsung's 48L 3D V-NAND Flash

Kevin Gibb, Product Line Mar
TechInsights
4/6/2016 04:40 PM EDT
9 comments post a comment

Like 16 Tweet Shar

**The highly anticipated Samsu
memory is out in the market,
first look.**

Samsung had announced its 25
K9AFGY8S0M 3D V-NAND as
would be used in a variety of
on the market in early 2016. True to their word, we managed to fi
them in their 2 TB capacity, mSATA, T3 portable SSD shown in
Figure 1.

designlines WIRELESS & NETWORKING

Slideshow

Facebook Likes Intel's 3D XPoint

Google joins open hardware effort

Rick Merritt

NO RATINGS
LOGIN TO RATE

May 18, 2016

IBM Puts 3D XPoint on Notice with 3 Bits/Cell PCM Brea

Tiffany Trader

IBM scientists have broken new ground in the
change memory technology (PCM) that puts a
XPoint technology from Intel and Micron. IBM s

Original URL: http://www.theregister.co.uk/2013/11/01/hp_memristor_2018/

HP 100TB Memristor drives by 2018 – if you're lucky, admits tech titan

Universal memory slow in coming

By Chris Mellor

Posted in Storage, 1st November 2013 02:28 GMT

Blocks and Files HP has warned *El Reg* not to get its hopes up too high after the tech titan's CTO Martin Fink suggested StoreServ arrays could be packed with 100TB Memristor drives come 2018.

In five years, according to Fink, DRAM and NANI
of the technologies: process shrinks will come to
off a cliff as a side effect of reducing the size of e

The HP answer to this scaling wall is Memristor,
to have DRAM-like speed and better-than-NAND
in Las Vegas that Memristor devices will be ready
also showed off a Memristor wafer, adding that it

tom's HARDWARE PRODUCT REVIEWS NEWS DEALS FORUM

Samsung's 10-Year Plan Starts With 128TB QLC SSD, 960 Successor

22 COMMENTS

by Chris Ramseyer August 8, 2017 at 12:30 PM

designlines MEMORY

News & Analysis

3D NAND Flash at 2 Cents per GB

BeSang wants to lower barr

R. Colin Johnson

7/18/2016 07:10 PM EDT

14 comments

designlines MEMORY

News & Analysis

Samsung Debuts 3D XPoint Killer

3D NAND variant stakes out high-end SSDs

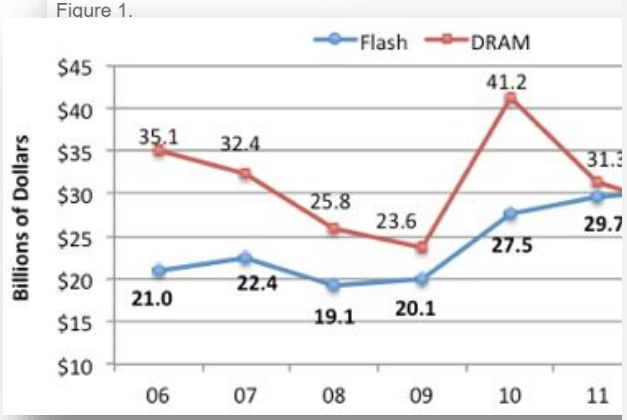
Rick Merritt

8/11/2016 00:01 AM EDT

5 comments

NO RATINGS
1 saves
LOGIN TO RATE

Like 56 Tweet Share 212 G+1 4



http://www.eetasia.com/STATIC/ARTICLE_IMAGES/201212/EEOL_2012DEC28_STOR_M

ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTUR

THE REVOLUTION IS HERE —

Intel at last announces Optane memory: DDR4 that never forgets

New memory offers huge capacities and persistence, but fits in a DDR4 slot.

PETER BRIGHT - 5/30/2018, 8:45 PM

Sandisk Ultra 400GB Micro SDXC UHS-I Card with Adapter - SDSQUAR-400G-GN6MA

by SanDisk

7,643 customer reviews
1000+ answered questions

Amazon's Choice for "400gb micro sd card"

List Price: $249.99
Deal of the Day: $99.99 prime
Ends in 07h 05m 08s
You Save: $150.00 (60%)

Capacity: 400GB

8GB	16GB	32GB
$11.99 prime	$7.79 Add-on Item	$9.79 prime
64GB	128GB	200GB
$15.99 prime	$26.71 prime	$34.99 prime
256GB	400GB	
$51.99 prime	$99.99 prime	

Forbes / Tech

JUL 28, 2015 @ 2:46 PM 7,391 VIEWS

Intel And Micron Jointly Announce Game-Changing 3D XPoint Memory Technology

DESIGNLINES | MEMORY DESIGNLINE

Memory Forecast to Accoun 53% of Semiconductor Cape

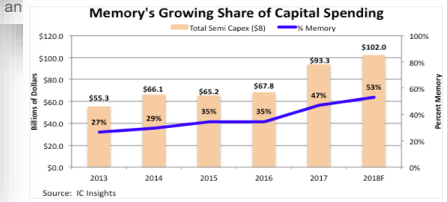
By Dylan McGrath, 08.29.18

Share Post Share on Facebook Share on Twitter G+

SAN FRANCISCO — Capital spending for memory chips is expected to accoun industry capex of $102 billion this year, nearly twice the percentage that mem five years ago, according to market research firm IC Insights.

With all NAND flash vendors ramping up 3D NAND capacity, NAND-related c are forecast to total more than $31 billion, 31% of the semiconductor industry the latest edition of IC Insights' McClean Report. The total for NAND capex w increase of 13% over 2017, when NAND flash capex grew by 91%.

Meanwhile, the report forecasts that capital spending for DRAM and SRAM wi any other industry segment, growing 41% in 2018 after an 82% increase last y capex is expected to total $22.9 billion, 22% of the industry-wide total, accord

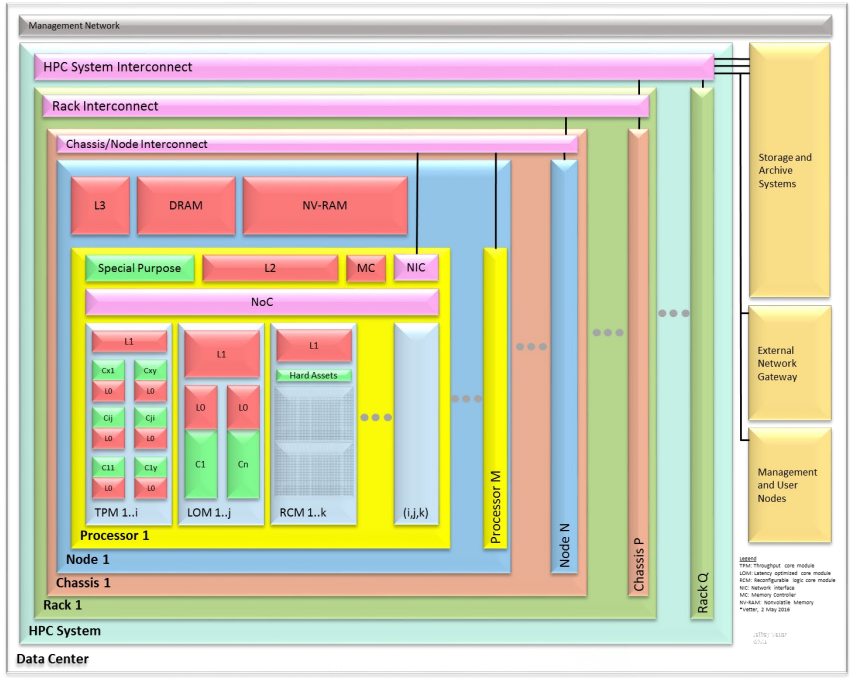


The forecasted total of $102 billion for the overall semiconductor industry — i upgrades to existing wafer fab lines and brand new manufacturing facilities —

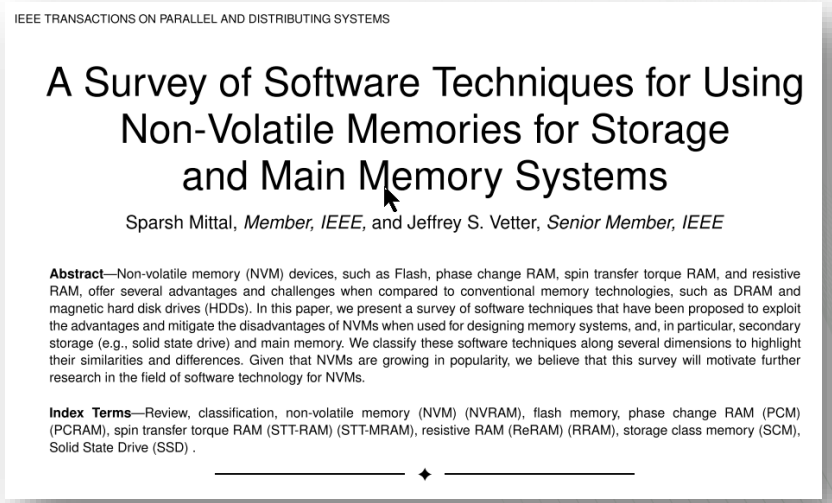
Programming NVM Systems Portably

NVM Design Choices

- Dimensions
  - Integration point
  - Exploit persistence
    - ACID?
  - Scalability
  - Programming model
- Our Approaches
  - Transparent access to NVM from GPU
  - NVL-C: expose NVM to user/applications
  - Papyrus: parallel aggregate persistent memory
  - Many others (See S. Mittal and J. S. Vetter, "A Survey of Software Techniques for Using Non-Volatile Memories for Storage and Main Memory Systems," in IEEE TPDS 27:5, pp. 1537-1550, 2016)



http://j.mp/nvm-sw-survey



IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTING SYSTEMS

A Survey of Software Techniques for Using Non-Volatile Memories for Storage and Main Memory Systems

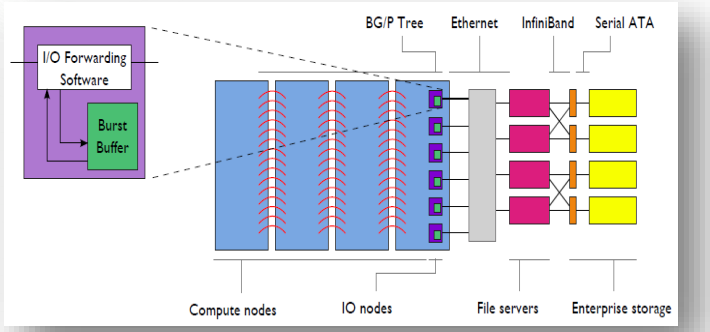
Sparsh Mittal, *Member, IEEE*, and Jeffrey S. Vetter, *Senior Member, IEEE*

Abstract—Non-volatile memory (NVM) devices, such as Flash, phase change RAM, spin transfer torque RAM, and resistive RAM, offer several advantages and challenges when compared to conventional memory technologies, such as DRAM and magnetic hard disk drives (HDDs). In this paper, we present a survey of software techniques that have been proposed to exploit the advantages and mitigate the disadvantages of NVMs when used for designing memory systems, and, in particular, secondary storage (e.g., solid state drive) and main memory. We classify these software techniques along several dimensions to highlight their similarities and differences. Given that NVMs are growing in popularity, we believe that this survey will motivate further research in the field of software technology for NVMs.

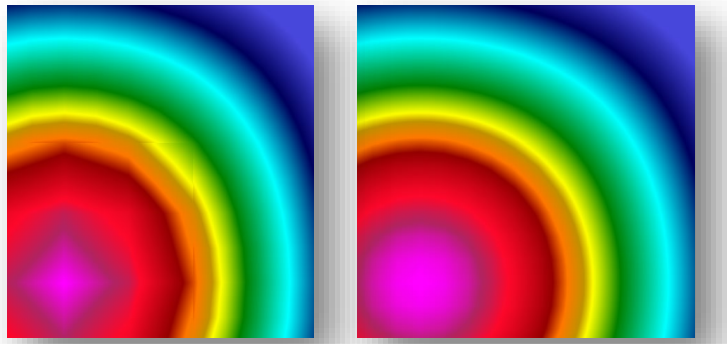
Index Terms—Review, classification, non-volatile memory (NVM) (NVRAM), flash memory, phase change RAM (PCM) (PCRAM), spin transfer torque RAM (STT-RAM) (STT-MRAM), resistive RAM (ReRAM) (RRAM), storage class memory (SCM), Solid State Drive (SSD) .

NVM Opportunities in Applications

- Burst Buffers, C/R [Liu, et al., MSST 2012]



- In situ visualization and analytics



http://ft.ornl.gov/eavl

- Persistent data structures like materials tables

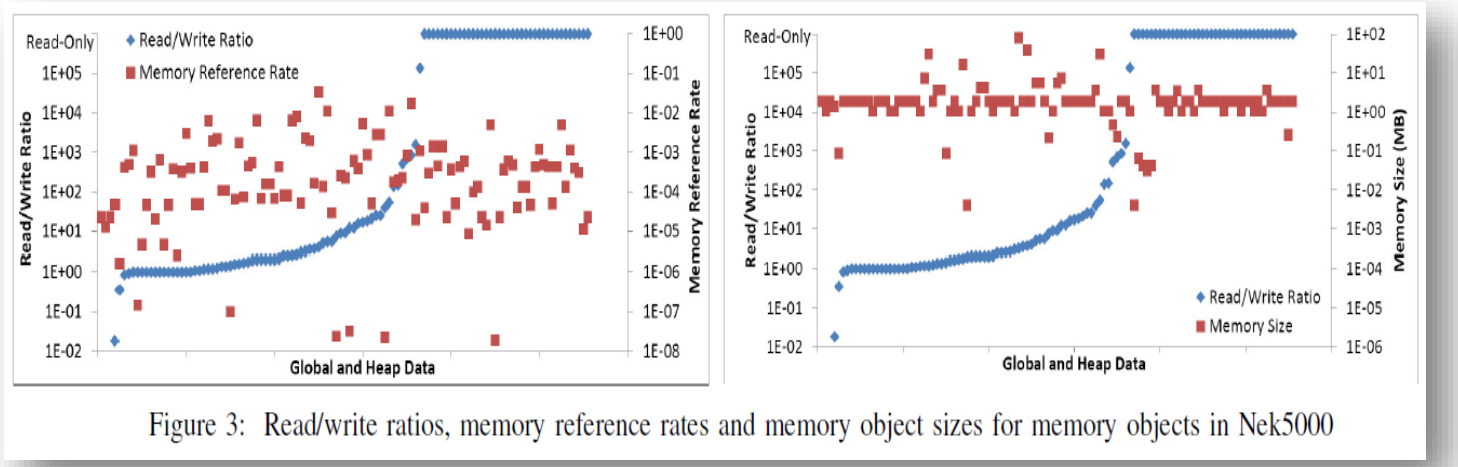


Figure 3: Read/write ratios, memory reference rates and memory object sizes for memory objects in Nek5000

Empirical results show many reasons…

- Lookup, index, and permutation tables
- Inverted and 'element-lagged' mass matrices
- Geometry arrays for grids
- Thermal conductivity for soils
- Strain and conductivity rates
- Boundary condition data
- Constants for transforms, interpolation
- MC Tally tables, cross-section materials tables…

OAK RIDGE National Laboratory

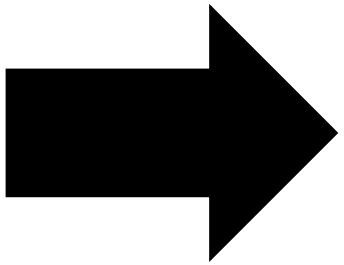
J.S. Vetter and S. Mittal, "Opportunities for Nonvolatile Memory Systems in Extreme-Scale High-Performance Computing," *Computing in Science & Engineering, 17(2):73-82, 2015.*

Transparent Runtime Support for NVM from GPUs

DRAGON: API and Integration

Out-of-Core using CUDA

```
// Allocate host & device memory
h_buf = malloc(size);
cudaMalloc(&g_buf, size);

while() { // go over all chunks
    // Read-in data
    f = fopen(filepath, “r”);
    fread(h_buf, size, 1, f);

    // H2D Transfer
    cudaMemcpy(g_buf, h_buf, H2D);

    // GPU compute
    compute_on_gpu(g_buf);

    // Transfer back to host
    cudaMemcpy(h_buf, g_buf, D2H);
    compute_on_host(h_buf);

    // Write out result
    fwrite(h_buf, size, 1, f);
}
```



DRAGON

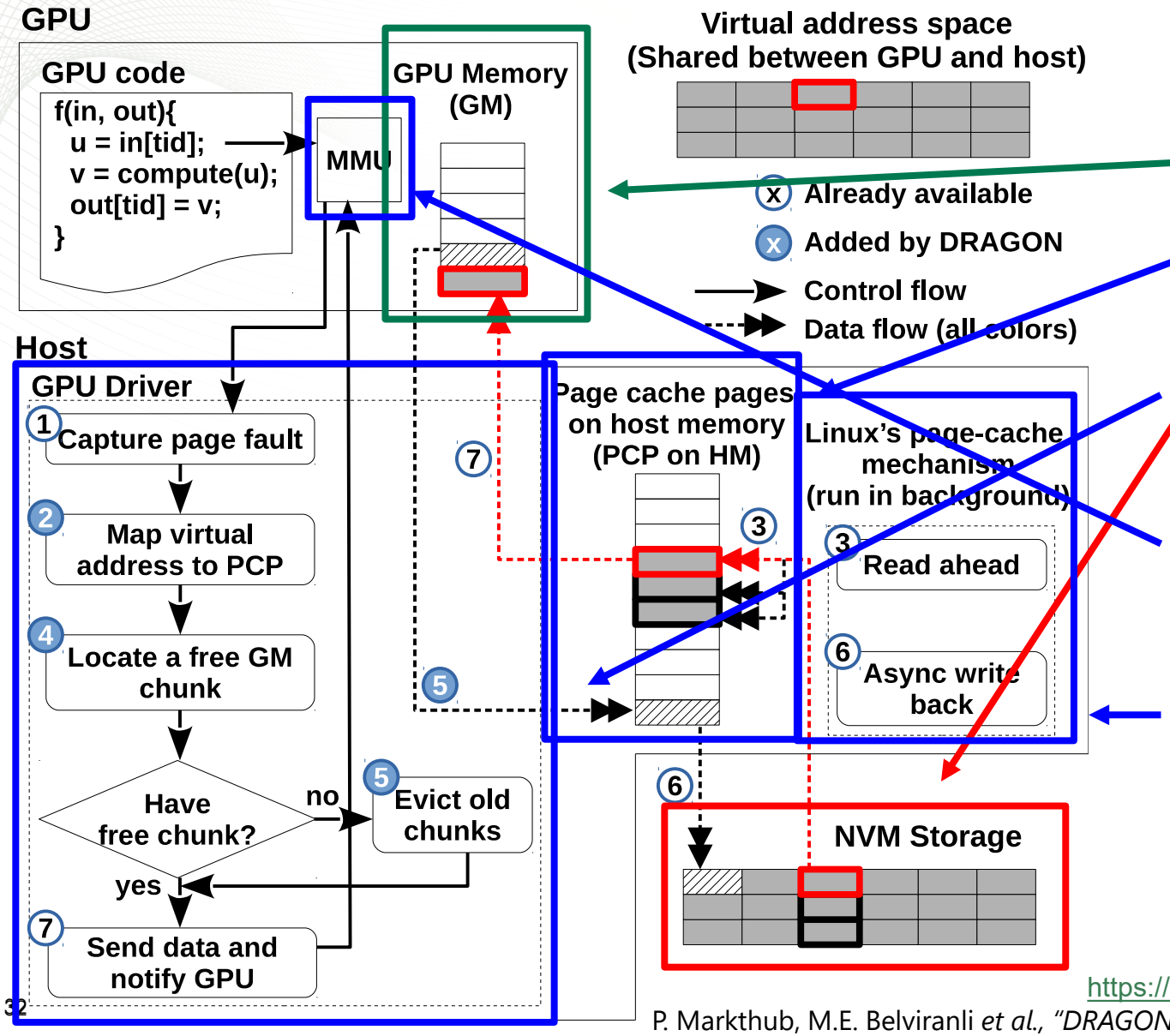
```
// mmap data to host and GPU
dragon_map(filepath, size,
    D_READ | D_WRITE, &g_buf);

// Accessible on both host and GPU
compute_on_gpu(g_buf);
compute_on_host(g_buf);

// Implicitly called when program
exits
dragon_sync(g_buf);
dragon_unmap(g_buf);
```

Notes

- Similar to NVIDIA’s Unified Memory (UM)
- Enable access to large memory on NVM
  - **UM is limited by host memory**

31

DRAGON Operations: Key Components



- **Three memory spaces:**
  - **GPU Mem (GM)** as 1st level cache
  - **Host Mem (HM)** as 2nd level cache
  - **NVM** as primary storage
- **Modified GPU driver**
  - Manage data movement & coherency
- **GPU MMU with HW Page Fault**
  - Manage GPU virtual memory mapping
- **Page cache**
  - Buffer & accelerate data access

https://github.com/pakmarkthub/dragon

P. Markthub, M.E. Belviranli *et al.*, *"DRAGON: Breaking GPU Memory Capacity Limits with Direct NVM Access," in SC18, 2018*

32

Results with Caffe

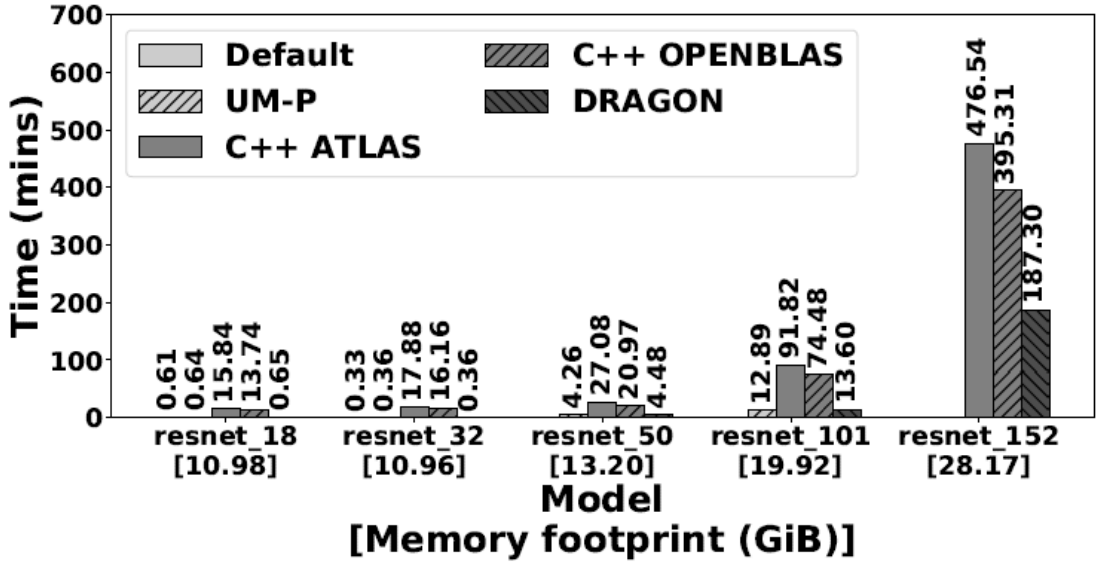


Figure 6: Comparison of ResNet execution times on Caffe.

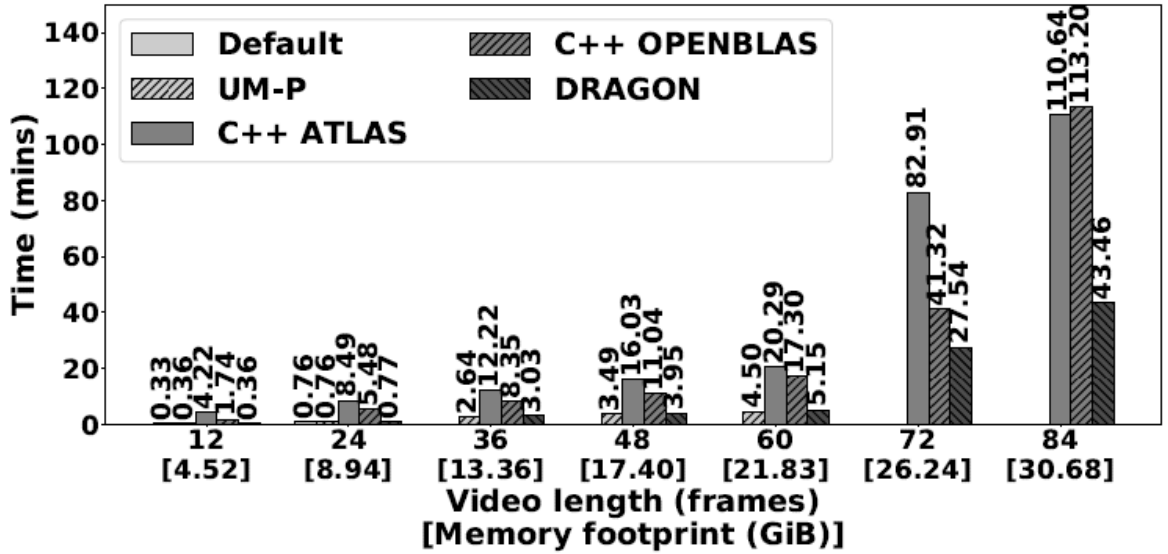


Figure 7: Comparison of C3D the execution times on Caffe.

- Improves capability and productivity
  - Larger problem sizes transparently
  - Handles irregularity easily
  - Surprising performance on applications

Language support for NVM:
NVL-C - extending C to support NVM

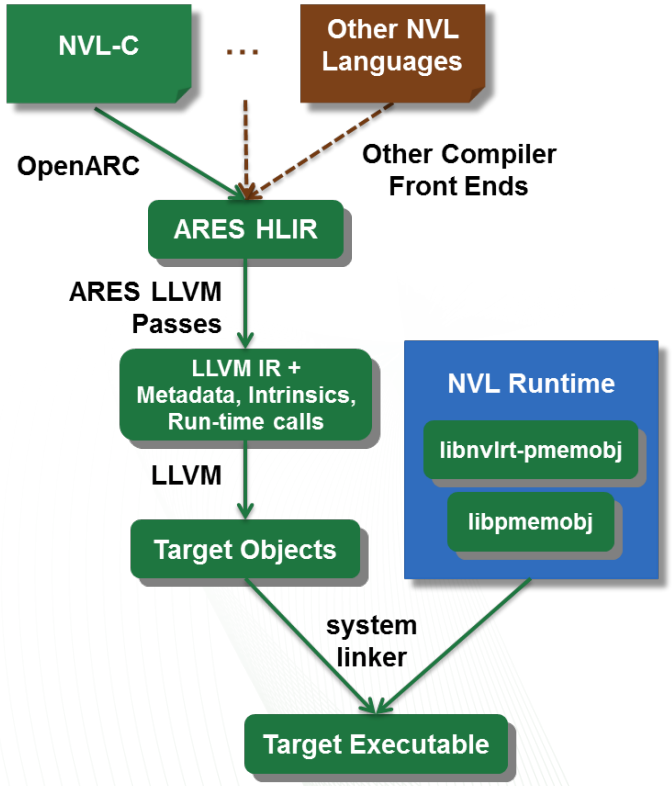
NVL-C: Portable Programming for NVMM

- Minimal, familiar, programming interface:
  - Minimal C language extensions.
  - App can still use DRAM.
- Pointer safety:
  - Persistence creates new categories of pointer bugs.
  - Best to enforce pointer safety constraints at compile time rather than run time.
- Transactions:
  - Prevent corruption of persistent memory in case of application or system failure.
- Language extensions enable:
  - Compile-time safety constraints.
  - NVM-related compiler analyses and optimizations.
- LLVM-based:
  - Core of compiler can be reused for other front ends and languages.
  - Can take advantage of LLVM ecosystem.

```
#include <nvl.h>
struct list {
  int value;
  nvl struct list *next;
};
void remove(int k) {
  nvl_heap_t *heap
    = nvl_open("foo.nvl");
  nvl struct list *a
    = nvl_get_root(heap, struct list);
  #pragma nvl atomic
  while (a->next != NULL) {
    if (a->next->value == k)
      a->next = a->next->next;
    else
      a = a->next;
  }
  nvl_close(heap);
}
```

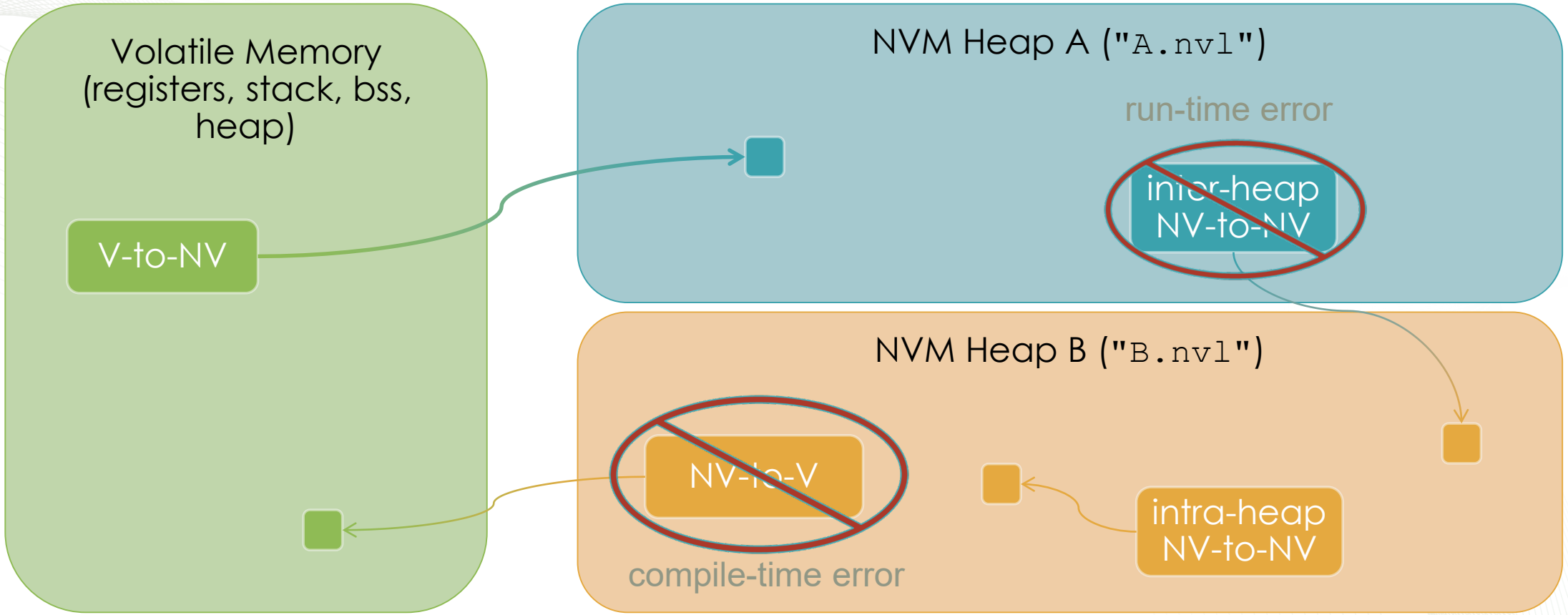
Pointer Class	Permitted
NV-to-V	no
V-to-NV	yes
intra-heap NV-to-NV	yes
inter-heap NV-to-NV	no

Table 1: Pointer Classes



J. Denny, S. Lee, and J.S. Vetter, "NVL-C: Static Analysis Techniques for Efficient, Correct Programming of Non-Volatile Main Memory Systems," in *ACM High Performance Distributed Computing (HPDC). Kyoto: ACM, 2016*

Programming Model: Pointer types (like Coburn et al.)



avoids dangling pointers when memory segments close

Programming Model: Transactions: MATMUL Example

```
#include <nvl.h>
void matmul(nvl float a[I][J],
            nvl float b[I][K],
            nvl float c[K][J],
            nvl int *i)
{
  for (; *i<I; ++*i) {
    for (int j=0; j<J; ++j) {
      float sum = 0.0;
      for (int k=0; k<K; ++k)
        sum += b[*i][k] * c[k][j];
      a[*i][j] = sum;
    }
  }
}
```

- Store `i` in NVM
- Caller initializes `*i` to `0` when allocated
- To recover after failure, `matmul` resumes at old `*i`
- Problem: failure might have occurred before all of `a[*i-1]` became durable in NVM due to buffering and caching

Programming Model: Transactions: MATMUL Example

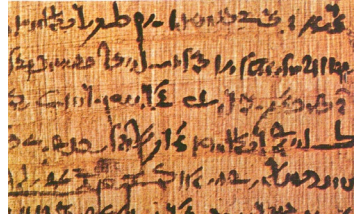
```
#include <nvl.h>
void matmul(nvl float a[I][J],
            nvl float b[I][K],
            nvl float c[K][J],
            nvl int *i)
{
  while (*i<I) {
    #pragma nvl atomic heap(heap)
    {
      for (int j=0; j<J; ++j) {
        float sum = 0.0;
        for (int k=0; k<K; ++k)
         sum += b[*i][k] * c[k][j];
        a[*i][j] = sum;
      }
      ++*i;
    }
  }
}
```

- **`nvl atomic`** pragma specifies explicit transaction that computes one row of `a`
- Transaction guarantees atomicity: both `*i` is incremented and one row of `a` is written durably, or neither
- Incomplete transaction rolled back after failure

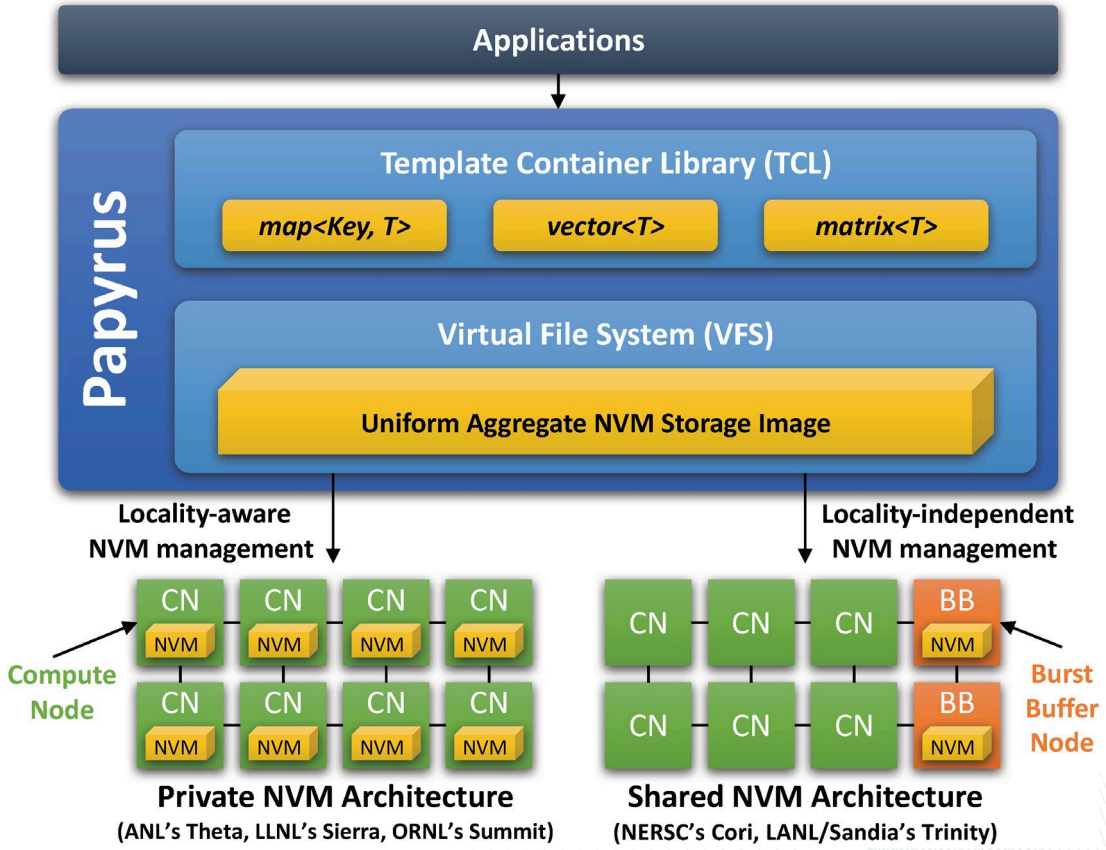
Programming Scalable NVM with Papyrus

Papyrus – Goals and Design

*Wikipedia: Papyrus can refer to a document written on sheets of papyrus, an early form of a book.



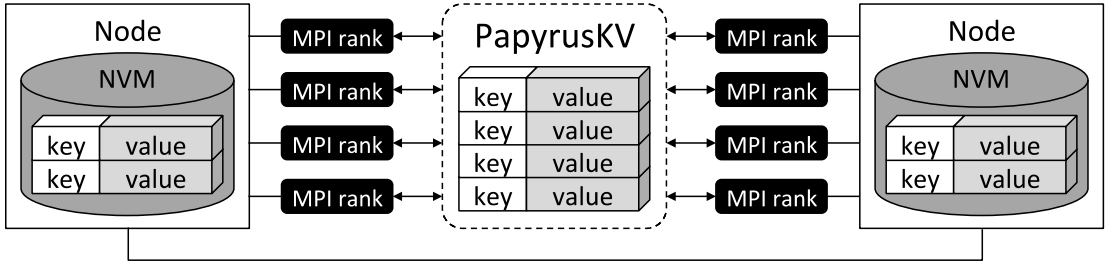
- Massive amounts of NVM in future systems will enable distributed persistent data structures – just say 'no' to I/O
- ***Papyrus*** is a novel programming system for aggregate NVM in the next generation HPC systems
  - ***P***arallel ***A***ggregate ***P***ersistent - ***YRU*** - ***S***torage
  - Portable and scalable programming interface
    - Private NVM & Shared NVM architectures
    - No centralized control
  - Papyrus Virtual File System
    - Interfaces to standard POSIX API
    - Allows for optimization on NVMe, Optane memory, etc.
  - Papyrus Template Container Library
    - C++ template container implementations



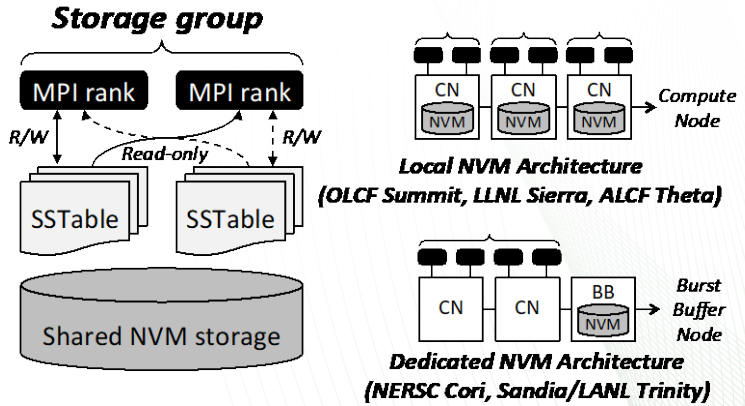
[1] J. Kim, S. Lee, and J.S. Vetter, "PapyrusKV: a high-performance parallel key-value store for distributed NVM architectures," in *SC17*.
[2] J. Kim, K. Sajjapongse, S. Lee, and J.S. Vetter, "Design and Implementation of Papyrus: Parallel Aggregate Persistent Storage," in IPDPS 2017.

PapyrusKV: A High-Performance Parallel Key-Value Store for Distributed NVM Architectures

- Leverage emerging NVM technologies
  - High performance
  - High capacity
  - Persistence property
- Designed for the next-generation DOE systems
  - Portable across local NVM and dedicated NVM architectures
  - An embedded key-value store (no system-level daemons and servers)
  - Scalability and performance
- Designed for HPC applications
  - MPI/UPC-interoperable
  - Application customizability
    - Memory consistency models (sequential and relaxed)
    - Protection attributes (read-only, write-only, read-write)
    - Load balancing
  - Zero-copy workflow, asynchronous checkpoint/restart



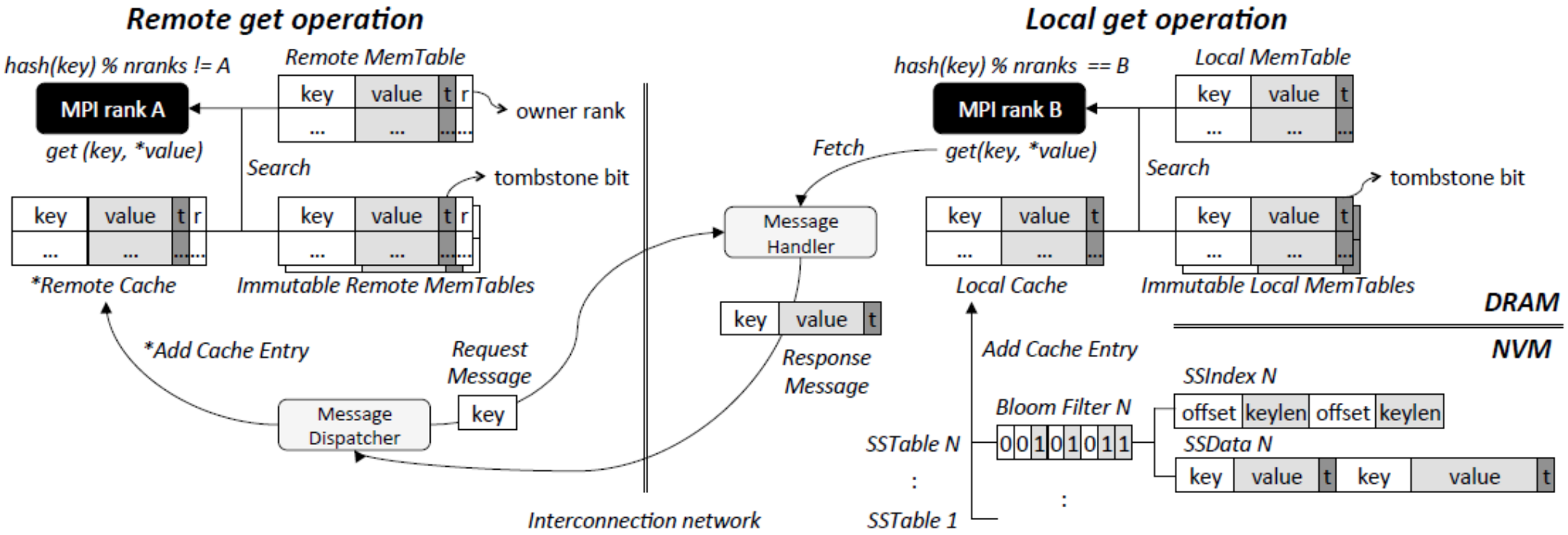
PapyrusKV stores keys and values in arbitrary byte arrays across multiple NVM devices in a distribute system



PapyrusKV is portable across local NVM and dedicated NVM architectures

J. Kim, S. Lee, and J. S. Vetter, "PapyrusKV: A High-Performance Parallel Key-Value Store for Distributed NVM Architectures," *In Proc. of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2017

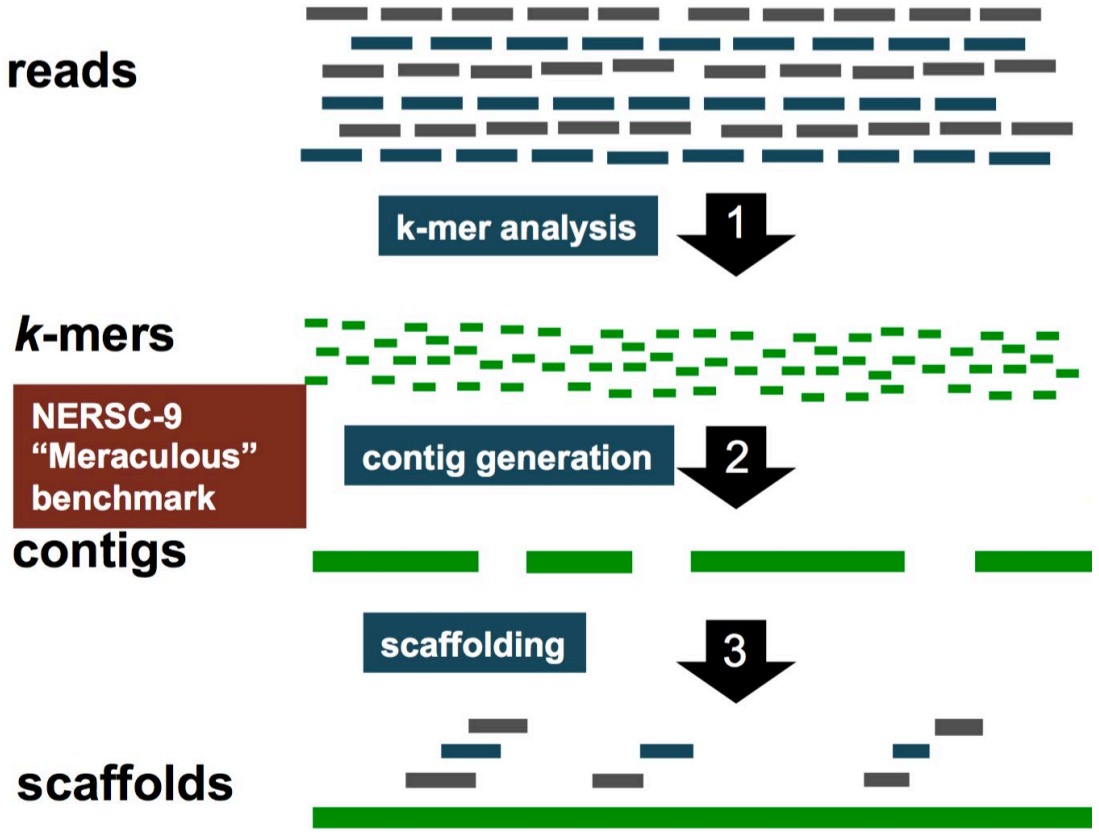
PapyrusKV Example Get operations



Present design allows remote cache only for RO data.

ECP Application Case Study 1 Meraculous (UPC)

- A parallel De Bruijin graph construction ar genome assembly
  - ExaBiome, Exascale Solutions for Microbiome



Graphic from ExaBiome: Exascale Solutions to Microbiome Analysis (LBNL, LANL, JGI), 2017

Table 1: Source lines of code.

Source file	UPC	UPC+PapyrusKV
meraculous.c	469	475 (+6)
buildUFXhashBinary.h	315	173 (-143)
kmer_hash.h	457	129 (-328)
UU_traversal_final.h	1754	1724 (-30)
Modified Total	2995	2501 (-494)
Grand Total	5971	5477 (-494)

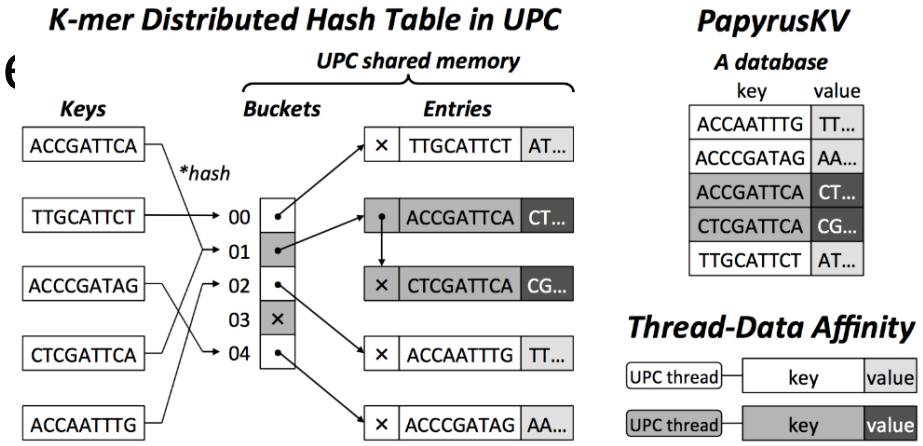


Figure 5: Distributed hash table implementations in UPC and PapyrusKV. *The same user *hash* function in the UPC application can be used in PapyrusKV.

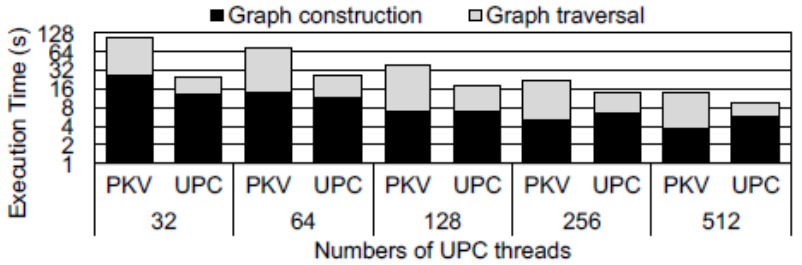


Figure 13: Meraculous performance comparison between PapyrusKV (PKV) and UPC on Cori.

NVM Implications

Implications

1. Device and architecture trends will have major impacts on HPC in coming decade
    1. NVM in HPC systems is real!
    2. Entirely possible to have an Exabyte of NVM in upcoming systems!
2. Performance trends of system components will create new opportunities and challenges
    1. Winners and losers
3. Sea of NVM allows/requires applications to operate differently
    1. Sea of NVM will permit applications to run for weeks without doing I/O to external storage system
    2. Applications will simply access local/remote NVM
    3. Longer term productive I/O will be ‘occasionally’ written to Lustre, GPFS
    4. Checkpointing (as we know it) will disappear
4. Requirements for system design will change
    1. Increase in byte-addressable memory-like message sizes and frequencies
    2. Reduced traditional IO demands
    3. KV traffic could have considerable impact – need more applications evidence
    4. Need changes to the operational mode of the system

Recap

- Recent trends in extreme-scale HPC paint an ambiguous future
- Complexity is the next major hurdle
  - Heterogeneous compute
  - Deep memory with NVM
- New software solutions
  - Programming
    - Memory
      - DRAGON
      - NVL-C
      - Papyrus
    - Heterogeneity
      - OpenACC->FPGAs
      - Clacc for LLVM
- These changes will have a substantial impact on both software and application design

- Visit us
  - We host interns and other visitors year round
- Jobs in FTG
  - Postdoctoral Research Associate in Computer Science
  - Software Engineer
  - Computer Scientist
  - Visit http://jobs.ornl.gov
- Contact me vetter@ornl.gov