# Preparing for Extreme Heterogeneity in High Performance Computing

Jeffrey S. Vetter

*With many contributions from FTG Group and Colleagues*

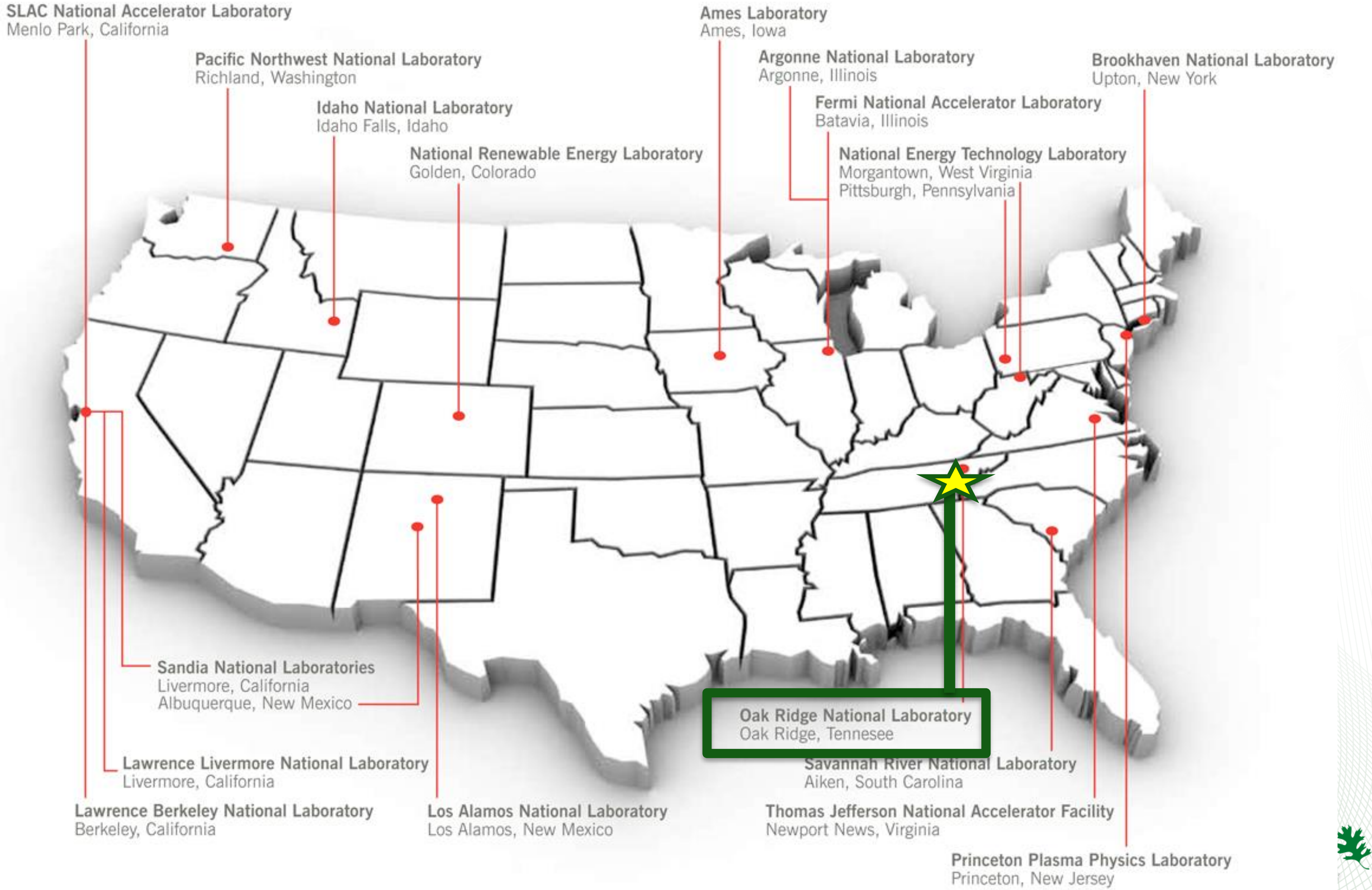
Barcelona Supercomputing Center
Technical University of Catalonia (UPC)
8 May 2019

ORNL is managed by UT-Battelle, LLC for the US 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)!!
  - Heterogeneous processor
    - OpenACC->FGPAs
    - Clacc – OpenACC support in LLVM (not covered today)
  - Emerging memory hierarchies (NVM)
    - DRAGON – transparent NVM access from GPUs
    - NVL-C – user management of nonvolatile memory in C
    - Papyrus – parallel aggregate persistent storage (not covered today)
- Performance prediction is critical for design and optimization (not covered today)

# Oak Ridge National Laboratory is the DOE Office of Science's Largest Lab



SLAC National Accelerator Laboratory
Menlo Park, California

Pacific Northwest National Laboratory
Richland, Washington

Idaho National Laboratory
Idaho Falls, Idaho

National Renewable Energy Laboratory
Golden, Colorado

Ames Laboratory
Ames, Iowa

Argonne National Laboratory
Argonne, Illinois

Fermi National Accelerator Laboratory
Batavia, Illinois

National Energy Technology Laboratory
Morgantown, West Virginia
Pittsburgh, Pennsylvania

Brookhaven National Laboratory
Upton, New York

Sandia National Laboratories
Livermore, California
Albuquerque, New Mexico

Lawrence Livermore National Laboratory
Livermore, California

Lawrence Berkeley National Laboratory
Berkeley, California

Los Alamos National Laboratory
Los Alamos, New Mexico

Oak Ridge National Laboratory
Oak Ridge, Tennesee

Savannah River National Laboratory
Aiken, South Carolina

Thomas Jefferson National Accelerator Facility
Newport News, Virginia

Princeton Plasma Physics Laboratory
Princeton, New Jersey

# Today, ORNL is a leading science and energy laboratory

$1.63B
FY18 expenditures

4,440
employees

$750M
modernization investment

Nation’s largest materials research portfolio

3,200
research guests annually

Nation’s most diverse energy portfolio

Forefront scientific computing facilities

World’s most intense neutron source

World-class research reactor

2,270
journal articles published in CY17

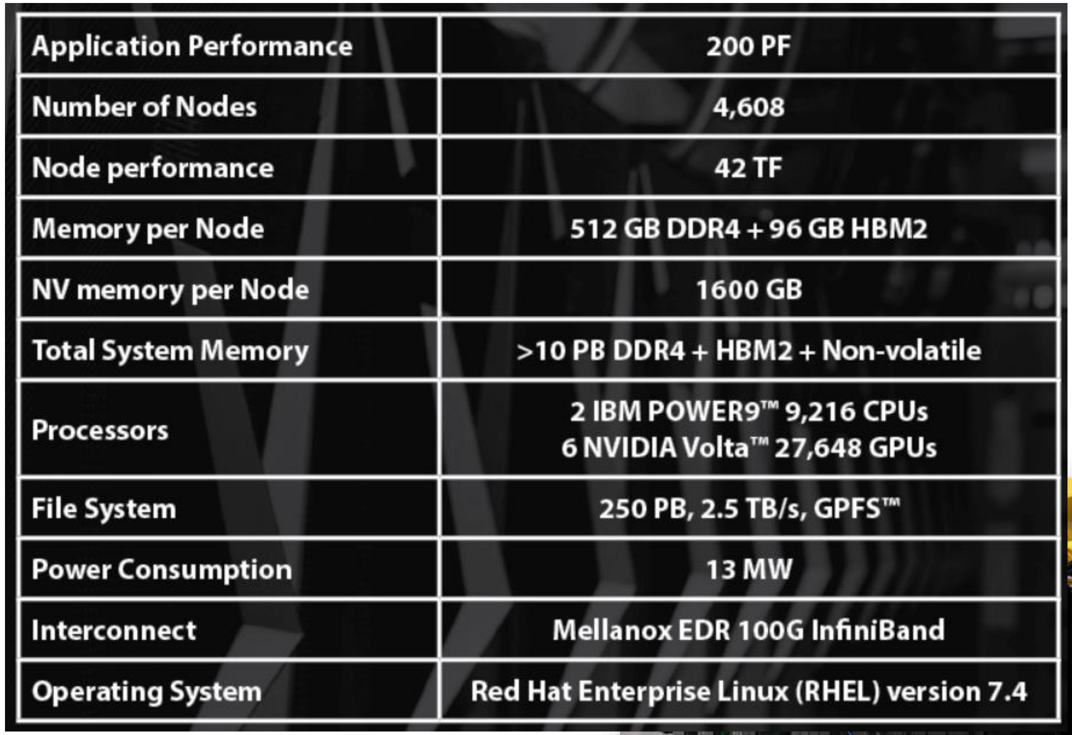
Managing major DOE projects: US ITER, exascale computing

258
invention disclosures in FY18

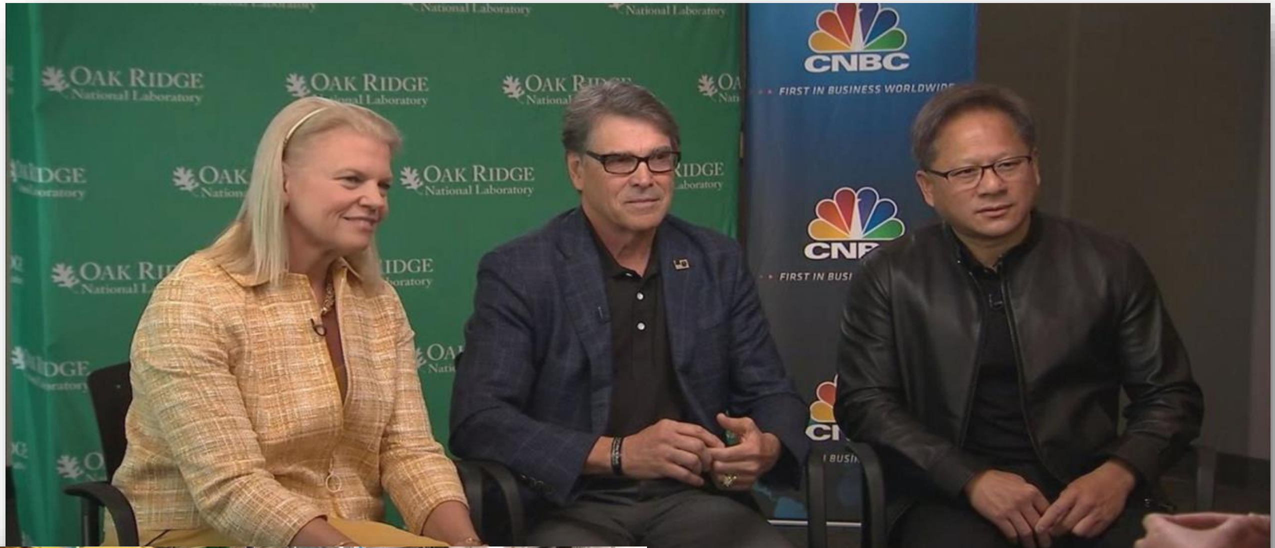
60
patents issued in FY18

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

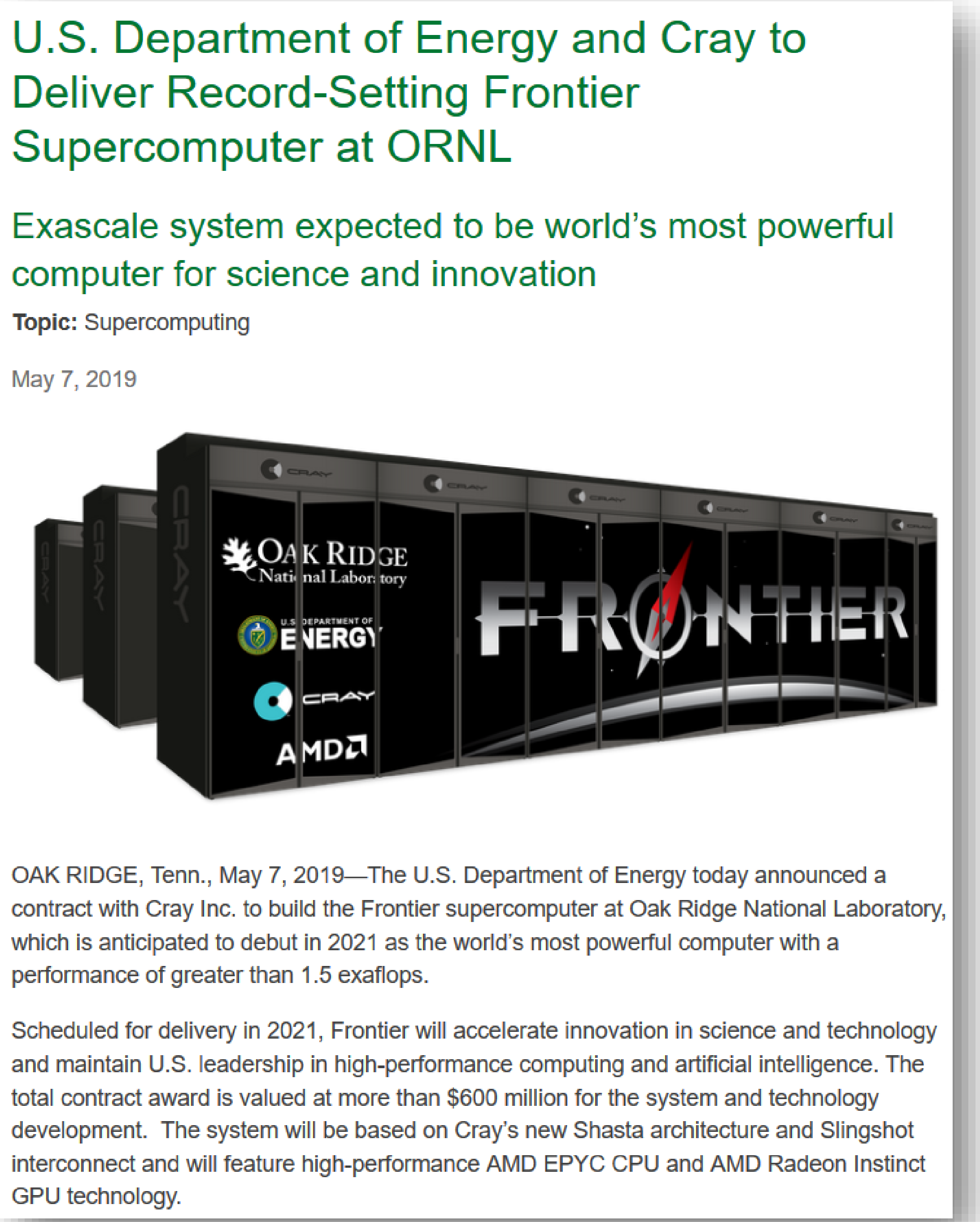
## #1 on Top 500



| 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<br>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



# U.S. Department of Energy and Cray to Deliver Record-Setting Frontier Supercomputer at ORNL

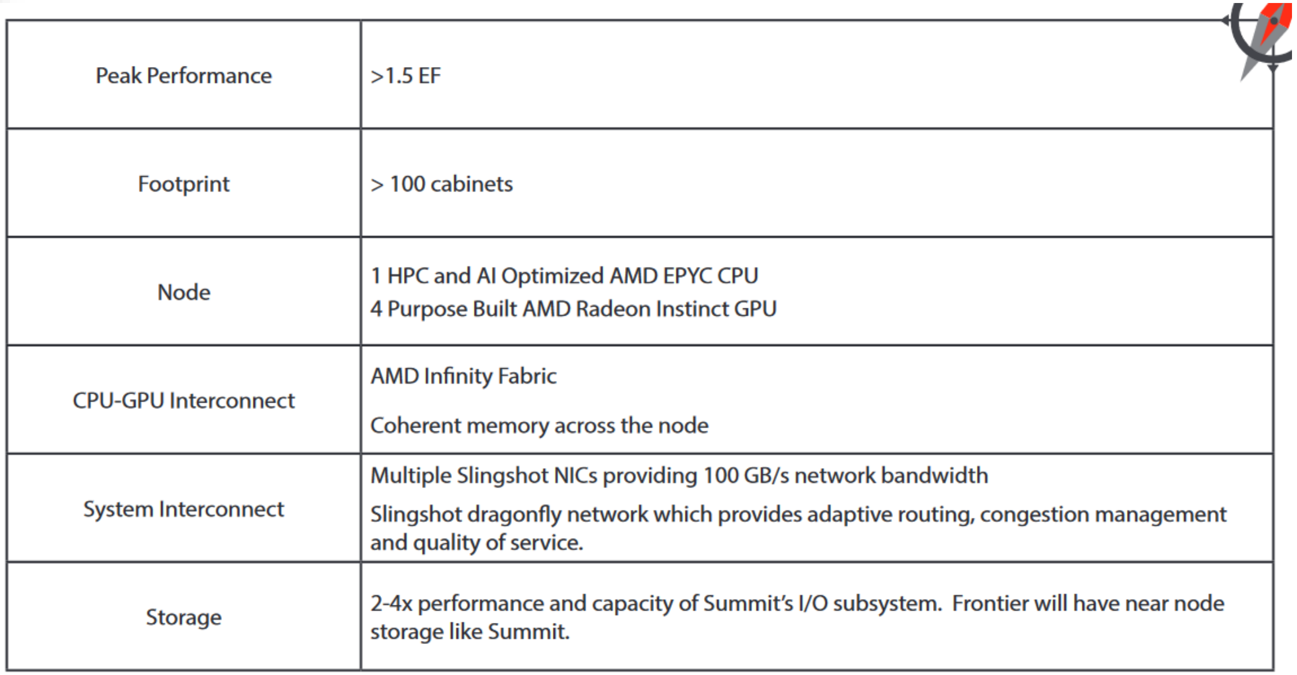
## Exascale system expected to be world's most powerful computer for science and innovation

**Topic:** Supercomputing

May 7, 2019

OAK RIDGE, Tenn., May 7, 2019—The U.S. Department of Energy today announced a contract with Cray Inc. to build the Frontier supercomputer at Oak Ridge National Laboratory, which is anticipated to debut in 2021 as the world's most powerful computer with a performance of greater than 1.5 exaflops.

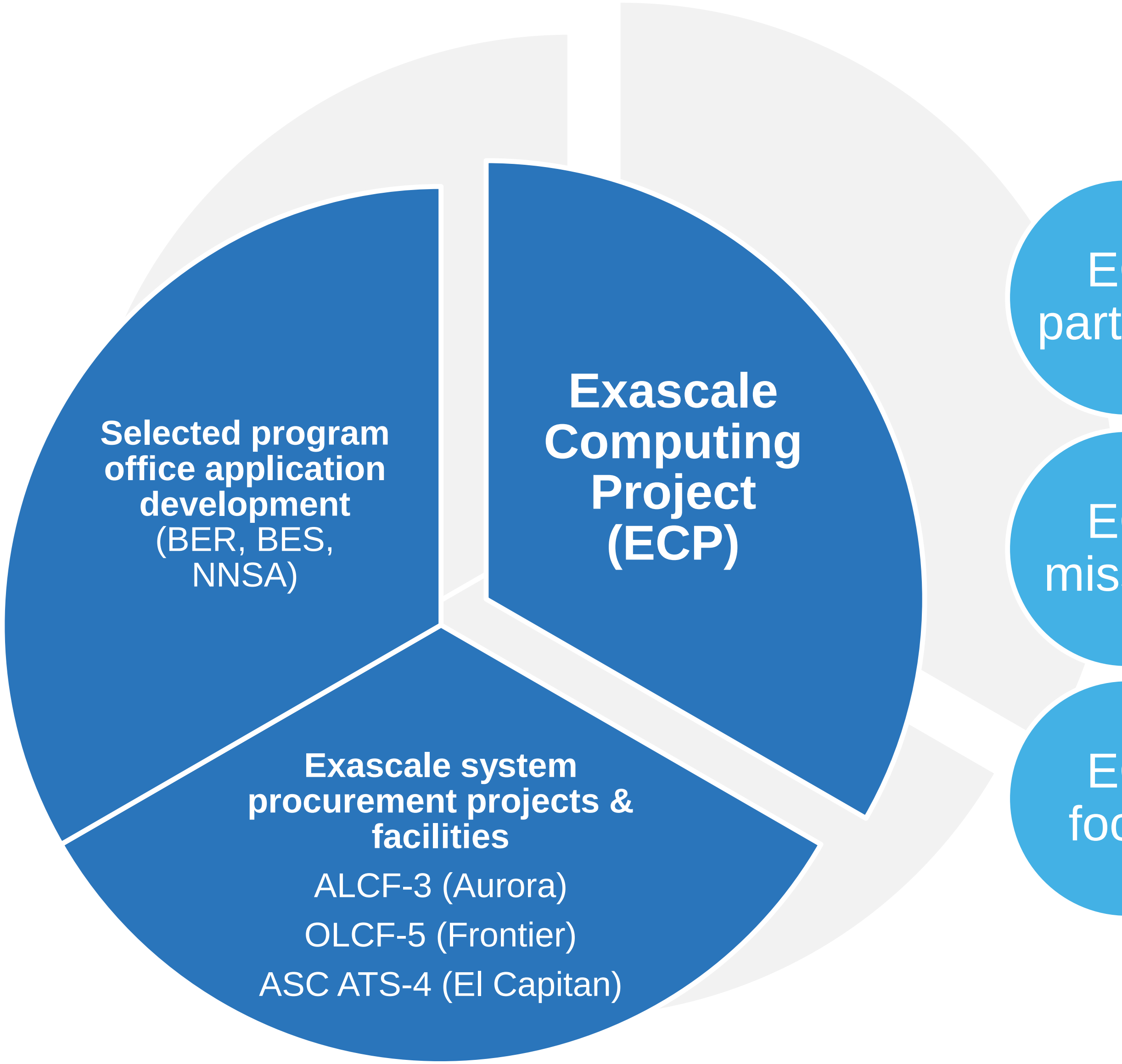
Scheduled for delivery in 2021, Frontier will accelerate innovation in science and technology and maintain U.S. leadership in high-performance computing and artificial intelligence. The total contract award is valued at more than $600 million for the system and technology development.  The system will be based on Cray's new Shasta architecture and Slingshot interconnect and will feature high-performance AMD EPYC CPU and AMD Radeon Instinct GPU technology.



| | |
|---|---|
| Peak Performance | >1.5 EF |
| Footprint | > 100 cabinets |
| Node | 1 HPC and AI Optimized AMD EPYC CPU<br>4 Purpose Built AMD Radeon Instinct GPU |
| CPU-GPU Interconnect | AMD Infinity Fabric<br>Coherent memory across the node |
| System Interconnect | Multiple Slingshot NICs providing 100 GB/s network bandwidth<br>Slingshot dragonfly network which provides adaptive routing, congestion management and quality of service. |
| Storage | 2-4x performance and capacity of Summit's I/O subsystem.  Frontier will have near node storage like Summit. |

# US Exascale Computing Project

# DOE Exascale Program: The Exascale Computing Initiative (ECI)



**Selected program office application development**
(BER, BES, NNSA)
**Exascale Computing Project (ECP)**
**Exascale system procurement projects & facilities**
ALCF-3 (Aurora)
OLCF-5 (Frontier)
ASC ATS-4 (El Capitan)

*Three Major Components of the ECI*

**ECI partners**: US DOE Office of Science (SC) and National Nuclear Security Administration (NNSA)

**ECI mission**: Accelerate R&D, acquisition, and deployment to deliver exascale computing capability to DOE national labs by the early- to mid-2020s

**ECI focus**: Delivery of an *enduring and capable exascale computing capability for use by a wide range of applications* of importance to DOE and the US

# ECP by the Numbers

7<br>YEARS

$1.7B

A seven-year, $1.7 B R&D effort that launched in 2016

6<br>CORE DOE<br>LABS

Six core DOE National Laboratories: Argonne, Lawrence Berkeley, Lawrence Livermore, Los Alamos, Oak Ridge, Sandia

- Staff from most of the 17 DOE national laboratories take part in the project

3<br>TECHNICAL<br>FOCUS<br>AREAS

Three technical focus areas (Application Development, Software Technology, Hardware and Integration) supported by project management expertise in the ECP Project Office

ECP<br>Project<br>Office

100<br>R&D TEAMS

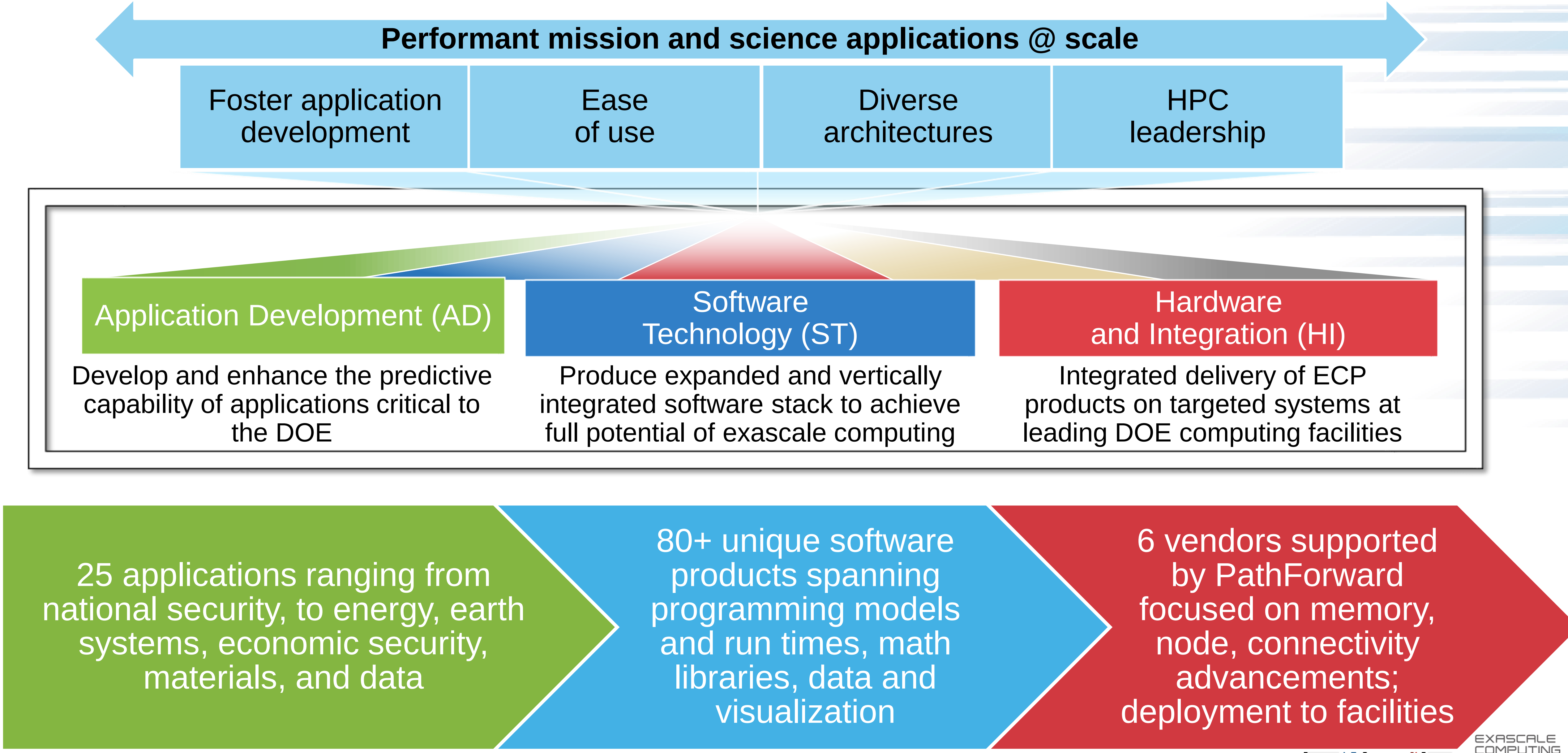
1000<br>RESEARCHERS

More than 100 top-notch R&D teams

Hundreds of consequential milestones delivered on schedule and within budget since project inception

EXASCALE<br>COMPUTING<br>PROJECT

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



**Performant mission and science applications @ scale**

Foster application development

Ease of use

Diverse architectures

HPC leadership

**Application Development (AD)**

Develop and enhance the predictive capability of applications critical to the DOE

**Software Technology (ST)**

Produce expanded and vertically integrated software stack to achieve full potential of exascale computing

**Hardware and Integration (HI)**

Integrated delivery of ECP products on targeted systems at leading DOE computing facilities

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

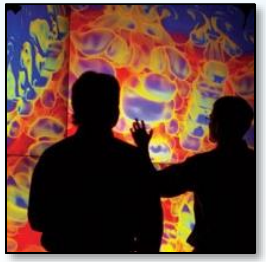
6 vendors supported by PathForward focused on memory, node, connectivity advancements; deployment to facilities

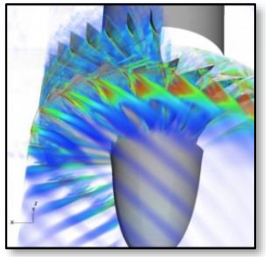
# ECP applications target national problems in 6 strategic areas

### 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

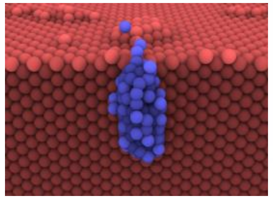
### Economic security

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

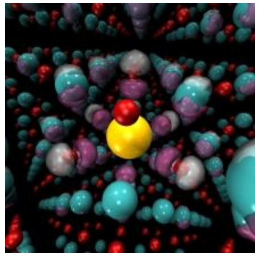
Cosmological probe of the standard model of particle physics

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

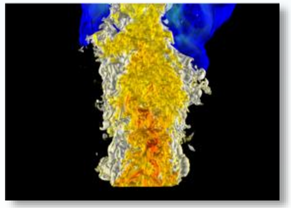
Accurate regional impact assessments in Earth system models

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

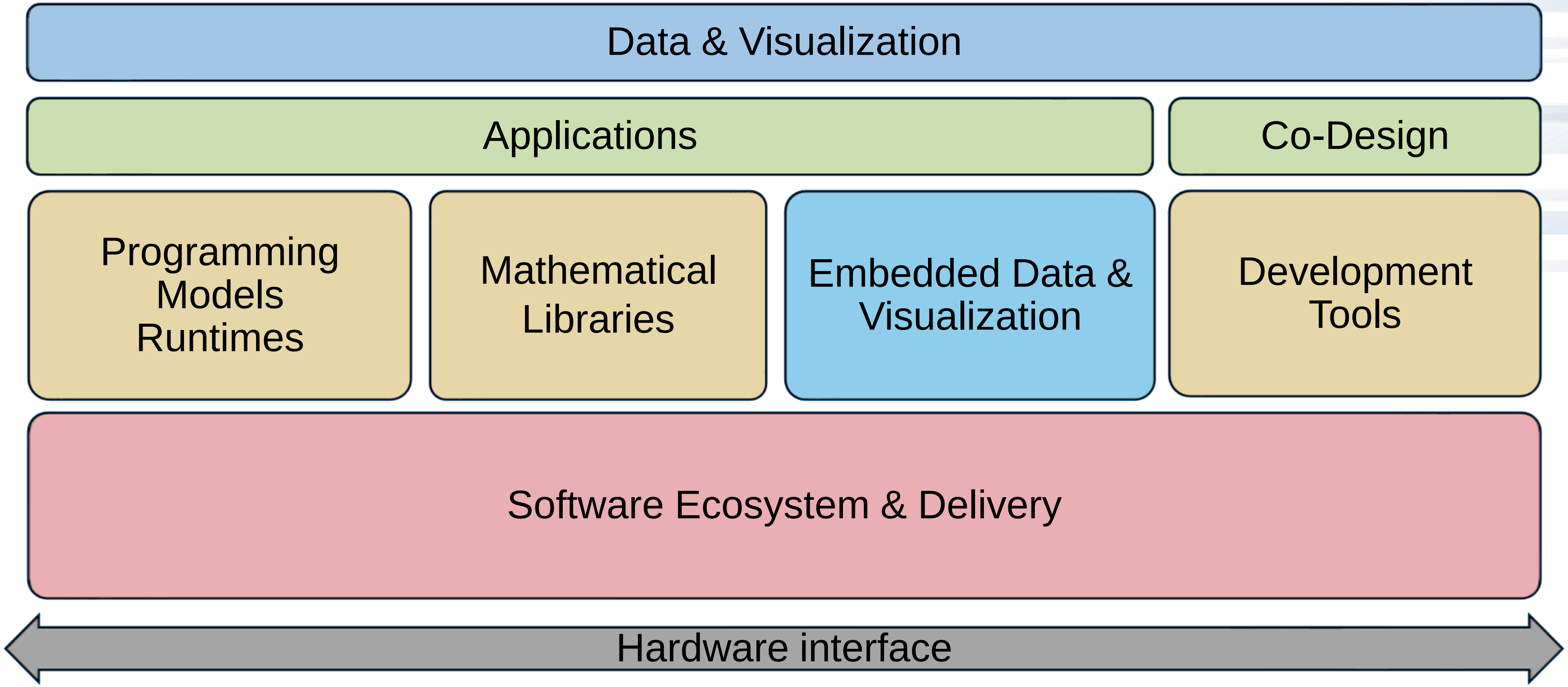
Metagenomics for analysis of biogeochemical cycles, climate change, environmental remediation

### Health care

Accelerate and translate cancer research



# ECP SW Stack: Strategic Alignment & Synergies



Data & Visualization

Applications

Co-Design

Programming Models Runtimes

Mathematical Libraries

Embedded Data & Visualization

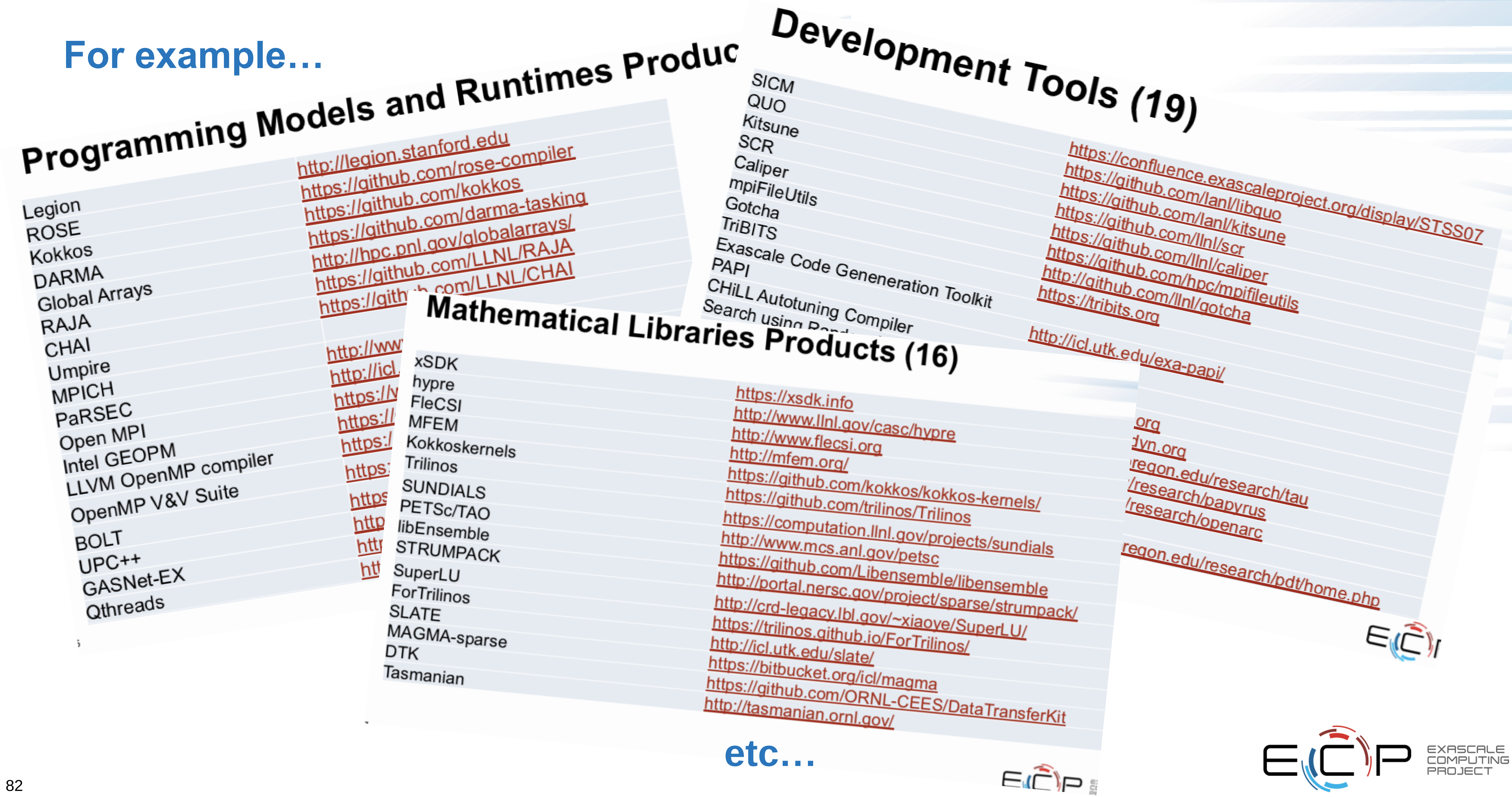
Development Tools

Software Ecosystem & Delivery

Hardware interface

# Many ECP ST products are available (many github)

For example…



## Programming Models and Runtimes Produc

| | |
|---|---|
| Legion | http://legion.stanford.edu |
| ROSE | https://github.com/rose-compiler |
| Kokkos | https://github.com/kokkos |
| DARMA | https://github.com/darma-tasking |
| Global Arrays | http://hpc.pnl.gov/globalarrays/ |
| RAJA | https://github.com/LLNL/RAJA |
| CHAI | https://github.com/LLNL/CHAI |
| Umpire | https://gith |

MPICH
PaRSEC
Open MPI
Intel GEOPM
LLVM OpenMP compiler
OpenMP V&V Suite
BOLT
UPC++
GASNet-EX
Qthreads

http://www
http://icl.
https://v
https://
https:/
https:
https
http
http
htt

## Development Tools (19)

| | |
|---|---|
| SICM | https://confluence.exascaleproject.org/display/STSS07 |
| QUO | https://github.com/lanl/libquo |
| Kitsune | https://github.com/lanl/kitsune |
| SCR | https://github.com/llnl/scr |
| Caliper | https://github.com/llnl/caliper |
| mpiFileUtils | http://github.com/hpc/mpifileutils |
| Gotcha | https://github.com/llnl/gotcha |
| TriBITS | https://tribits.org |
| Exascale Code Geneneration Toolkit | |
| PAPI | http://icl.utk.edu/exa-papi/ |
| CHiLL Autotuning Compiler | |
| Search using Rand | |

org
dyn.org
regon.edu/research/tau
/research/papyrus
/research/openarc
regon.edu/research/pdt/home.php

## Mathematical Libraries Products (16)

| | |
|---|---|
| xSDK | https://xsdk.info |
| hypre | http://www.llnl.gov/casc/hypre |
| FleCSI | http://www.flecsi.org |
| MFEM | http://mfem.org/ |
| Kokkoskernels | https://github.com/kokkos/kokkos-kernels/ |
| Trilinos | https://github.com/trilinos/Trilinos |
| SUNDIALS | https://computation.llnl.gov/projects/sundials |
| PETSc/TAO | http://www.mcs.anl.gov/petsc |
| libEnsemble | https://github.com/Libensemble/libensemble |
| STRUMPACK | http://portal.nersc.gov/project/sparse/strumpack/ |
| SuperLU | http://crd-legacy.lbl.gov/~xiaoye/SuperLU/ |
| ForTrilinos | https://trilinos.github.io/ForTrilinos/ |
| SLATE | http://icl.utk.edu/slate/ |
| MAGMA-sparse | https://bitbucket.org/icl/magma |
| DTK | https://github.com/ORNL-CEES/DataTransferKit |
| Tasmanian | http://tasmanian.ornl.gov/ |

etc…

# Software Development Kits (SDKs): A Key ST Design Feature

An important delivery vehicle for software products with a direct line of sight to ECP applications

## ECP software projects

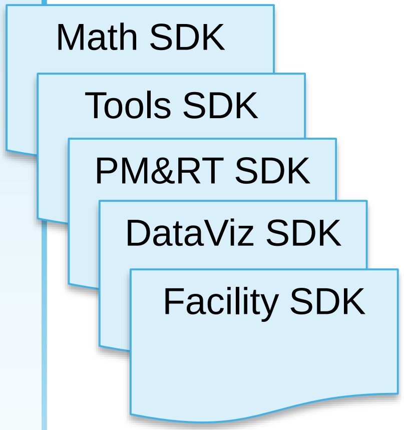
Each project to define (at least 2) release vectors

More projects ⟷ Fewer projects

### SDKs

Reusable software libraries embedded in applications; cohesive/interdependent libraries released as sets modeled on xSDK

- Regular coordinated releases
- Hierarchical collection built on Spack
- Products may belong to >1 SDK based on dependences
- Establish community policies for library development
- Apply Continuous Integration and other robust testing practices



### OpenHPC

Potential exit strategy for binary distributions

- Target similar software to existing OpenHPC stack
- Develop super-scalable release targeting higher end systems

### Direct2Facility

Platform-specific software in support of a specified 2021–2023 exascale system

- Software **exclusively** supporting a specific platform
- System software, some tools and runtimes

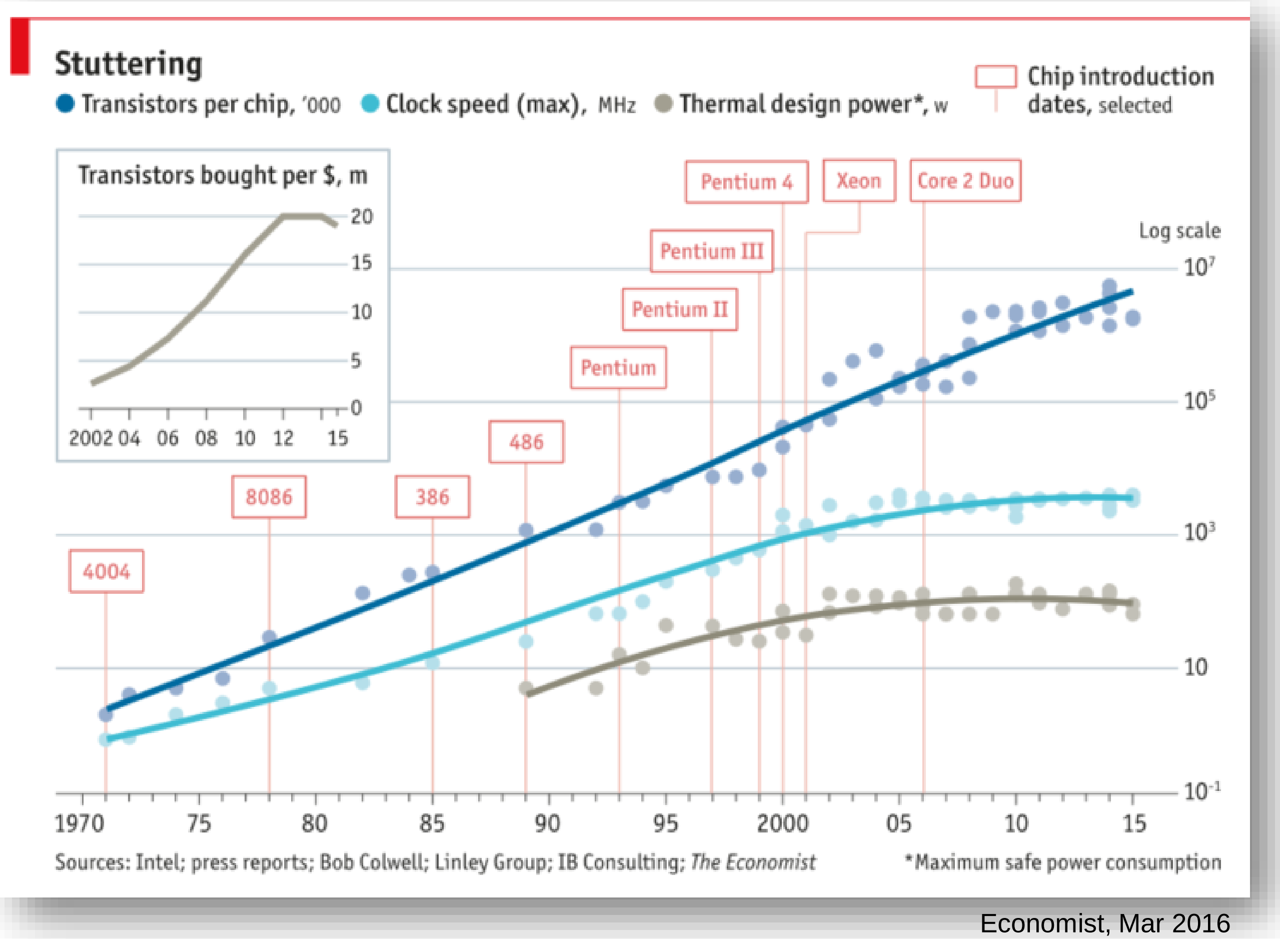
Assume all releases are delivered as “build from source” via Spack – at least initially

Focus on ensuring that software compiles robustly on all platforms of interest to ECP (including testbeds)

http://e4s.io

# Major Trends in Computing

# Contemporary devices are approaching fundamental limits



Sources: Intel; press reports; Bob Colwell; Linley Group; IB Consulting; *The Economist*

*Maximum safe power consumption

Economist, Mar 2016

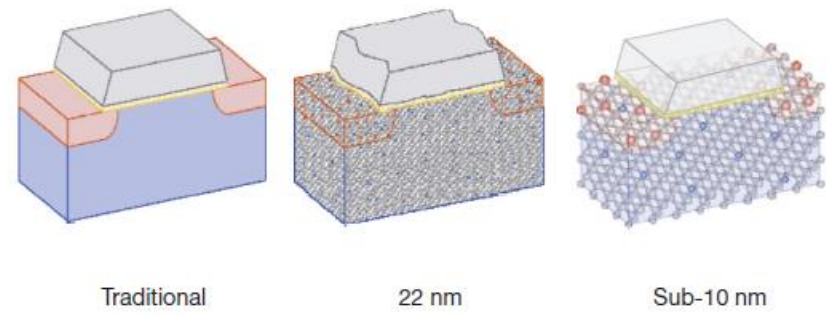


**Figure 1** | As a metal oxide–semiconductor field effect transistor (MOSFET) shrinks, the gate dielectric (yellow) thickness approaches several atoms (0.5 nm at the 22-nm technology node). Atomic spacing limits the

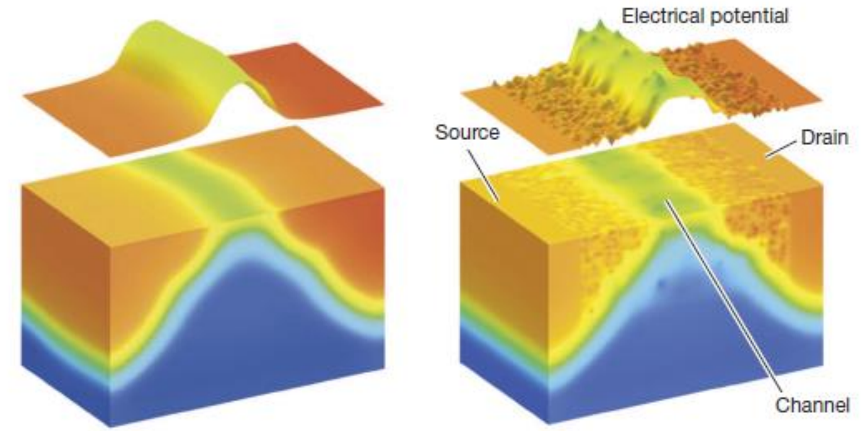


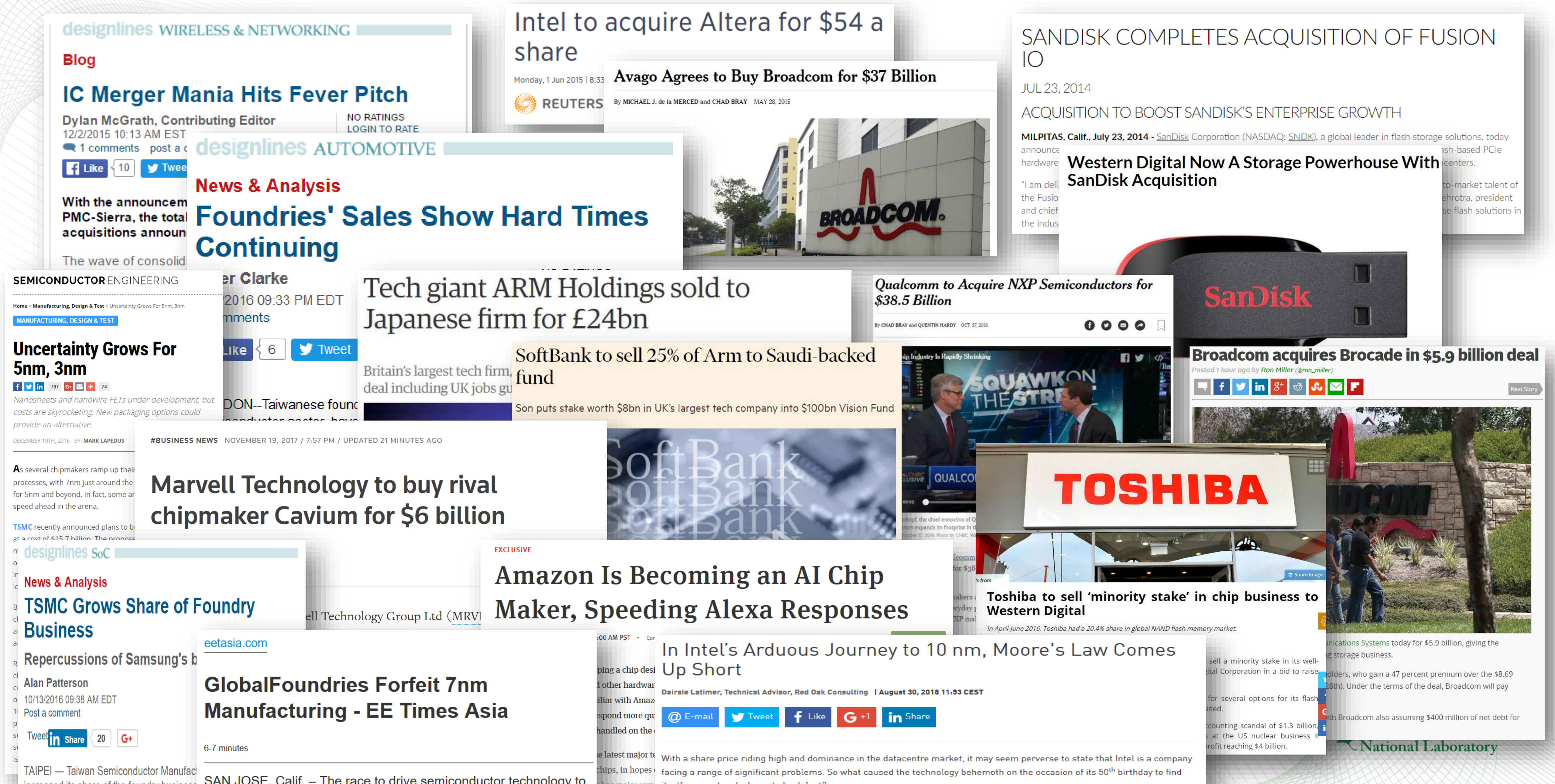
**Figure 2** | As a MOSFET transistor shrinks, the shape of its electric field departs from basic rectilinear models, and the level curves become disconnected. Atomic-level manufacturing variations, especially for dopant

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 … sh-based PCIe hardware … centers.

"I am deli … to-market talent of the Fusio … ehrotra, president and chief … se flash solutions in the indus

## Western Digital Now A Storage Powerhouse With SanDisk Acquisition

designlines AUTOMOTIVE

News & Analysis

## Foundries' Sales Show Hard Times Continuing

er Clarke
/2016 09:33 PM EDT
mments

Like 6 Tweet

DON--Taiwanese found

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 propose

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

Britain's largest tech firm
deal including UK jobs gu

## 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

## Qualcomm to Acquire NXP Semiconductors for $38.5 Billion

By CHAD BRAY and QUENTIN HARDY OCT. 27, 2016

hip Industry Is Rapidly Shrinking

nkopf, the chief executive of Q
ctors expands its footprint in th
October 27, 2016. Photo by CNBC W

lcomm
for $38

akers
ryday
XP mal

## Broadcom acquires Brocade in $5.9 billion deal

Posted 1 hour ago by Ron Miller (@ron_miller)

Next Story

nications Systems today for $5.9 billion, giving the
g storage business.

olders, who gain a 47 percent premium over the $8.69
28th). Under the terms of the deal, Broadcom will pay

ith Broadcom also assuming $400 million of net debt for

National Laboratory

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

## Marvell Technology to buy rival chipmaker Cavium for $6 billion

ell Technology Group Ltd (MRV

## Toshiba to sell 'minority stake' in chip business to Western Digital

*In April-June 2016, Toshiba had a 20.4% share in global NAND flash memory market.*

sell a minority stake in its well-
ital Corporation in a bid to raise

for several options for its flash
ided.

ccounting scandal of $1.3 billion,
at the US nuclear business it
rofit reaching $4 billion.

designlines SoC

News & Analysis

## TSMC Grows Share of Foundry Business

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

EXCLUSIVE

## Amazon Is Becoming an AI Chip Maker, Speeding Alexa Responses

:00 AM PST

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

eetasia.com

## GlobalFoundries Forfeit 7nm Manufacturing - EE Times Asia

6-7 minutes

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

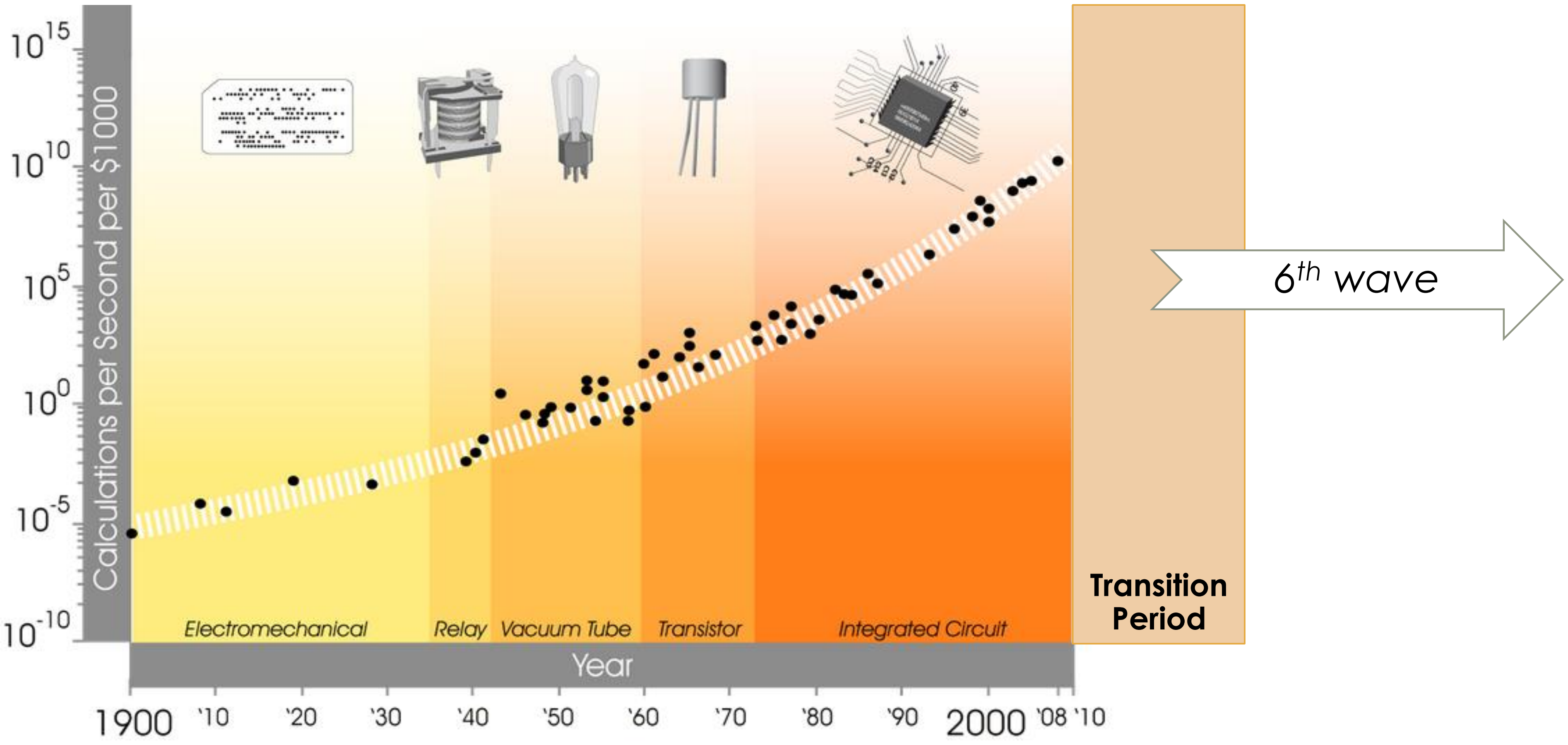
## In Intel's Arduous Journey to 10 nm, Moore's Law Comes Up Short

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

E-mail Tweet Like +1 Share

With a share price riding high and dominance in the datacentre market, it may seem perverse to state that Intel is a company facing a range of significant problems. So what caused the technology behemoth on the occasion of its 50th birthday to find

# Sixth Wave of Computing



Transition Period

6th wave

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

# Predictions for Transition Period

## 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

# 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

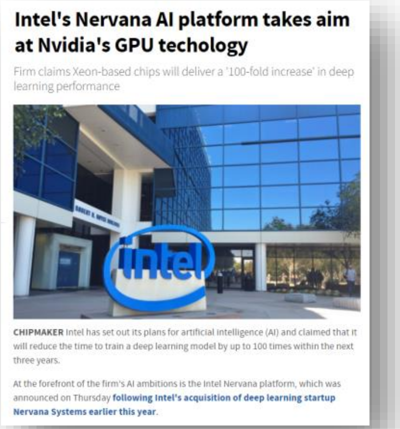
- Use CMOS more efficiently for our workloads
- Integrate components to boost performance and eliminate inefficiencies

## Emerging Technologies

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

# Pace of Architectural Specialization is Quickening

- Industry, 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



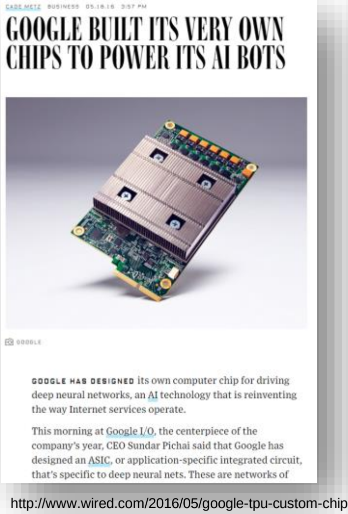
**Intel's Nervana AI platform takes aim at Nvidia's GPU techology**

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

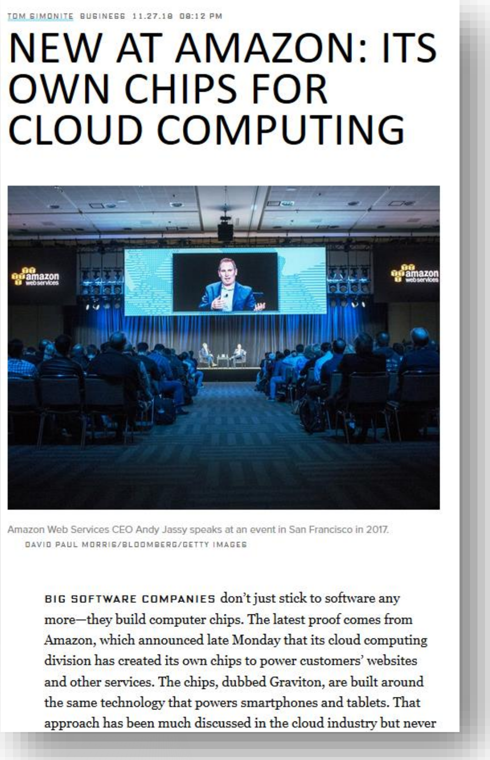
**GOOGLE BUILT ITS VERY OWN CHIPS TO POWER ITS AI BOTS**

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/



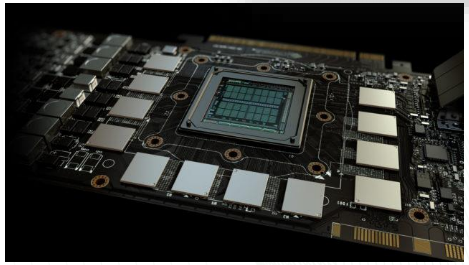
TOM SIMONITE BUSINESS 11.27.18 08:12 PM

**NEW AT AMAZON: ITS OWN CHIPS FOR CLOUD COMPUTING**

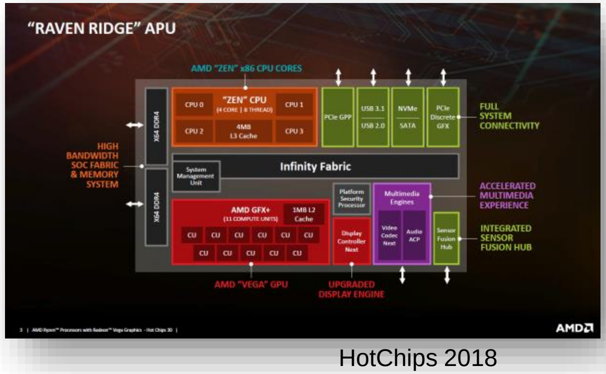
Amazon Web Services CEO Andy Jassy speaks at an event in San Francisco in 2017.
DAVID PAUL MORRIS/BLOOMBERG/GETTY IMAGES

BIG SOFTWARE COMPANIES don’t just stick to software any more—they build computer chips. The latest proof comes from Amazon, which announced late Monday that its cloud computing division has created its own chips to power customers’ websites and other services. The chips, dubbed Graviton, are built around the same technology that powers smartphones and tablets. That approach has been much discussed in the cloud industry but never

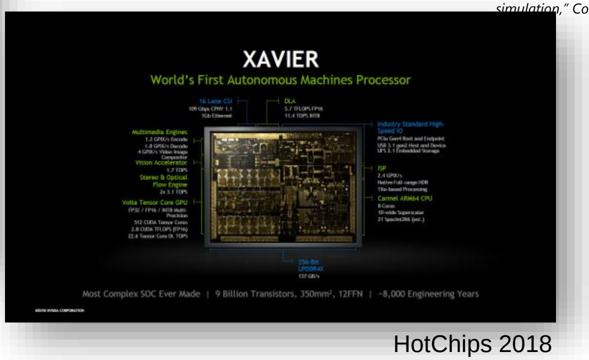
D.E. Shaw, M.M. Deneroff, R.O. Dror *et al., “Anton, a special-purpose machine for molecular dynamics simulation,” Communications of the ACM, 51(7):91-7, 2008.*



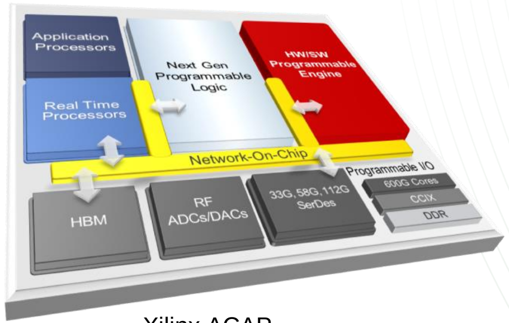
https://fossbytes.com/nvidia-volta-gddr6-2018/



HotChips 2018



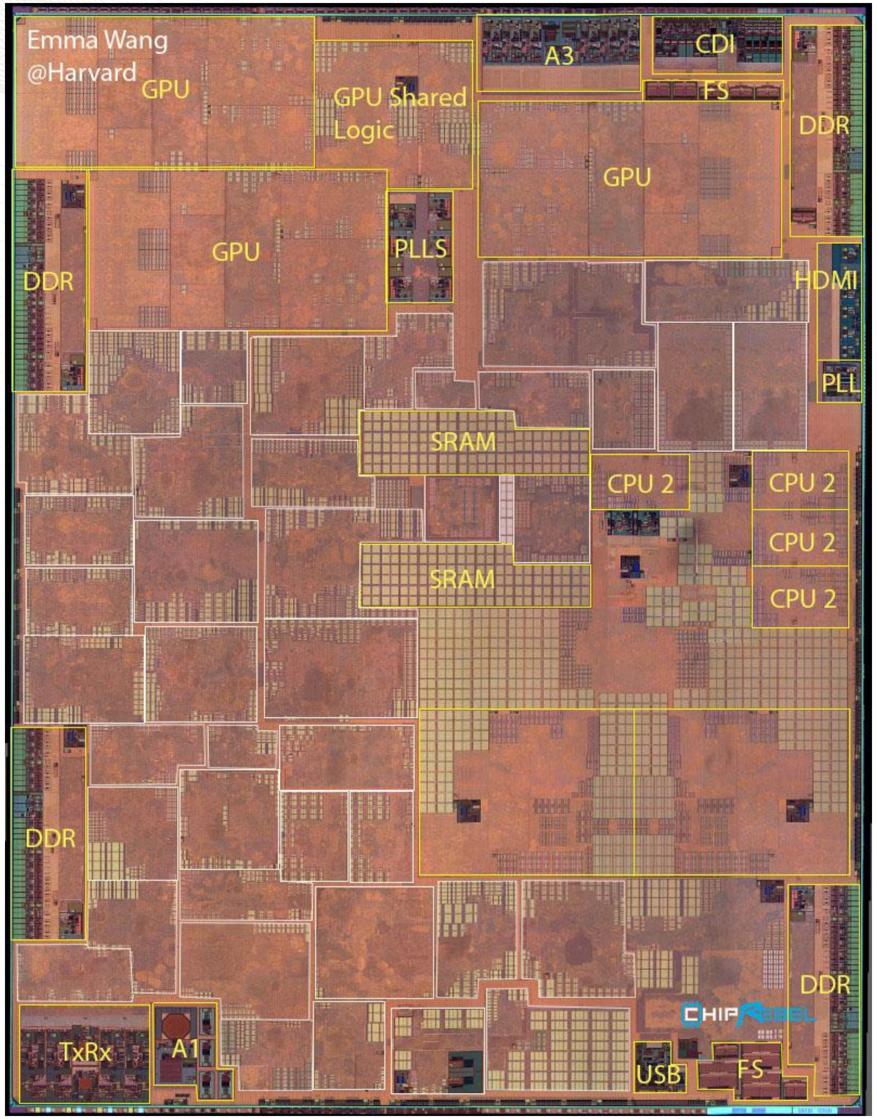
HotChips 2018

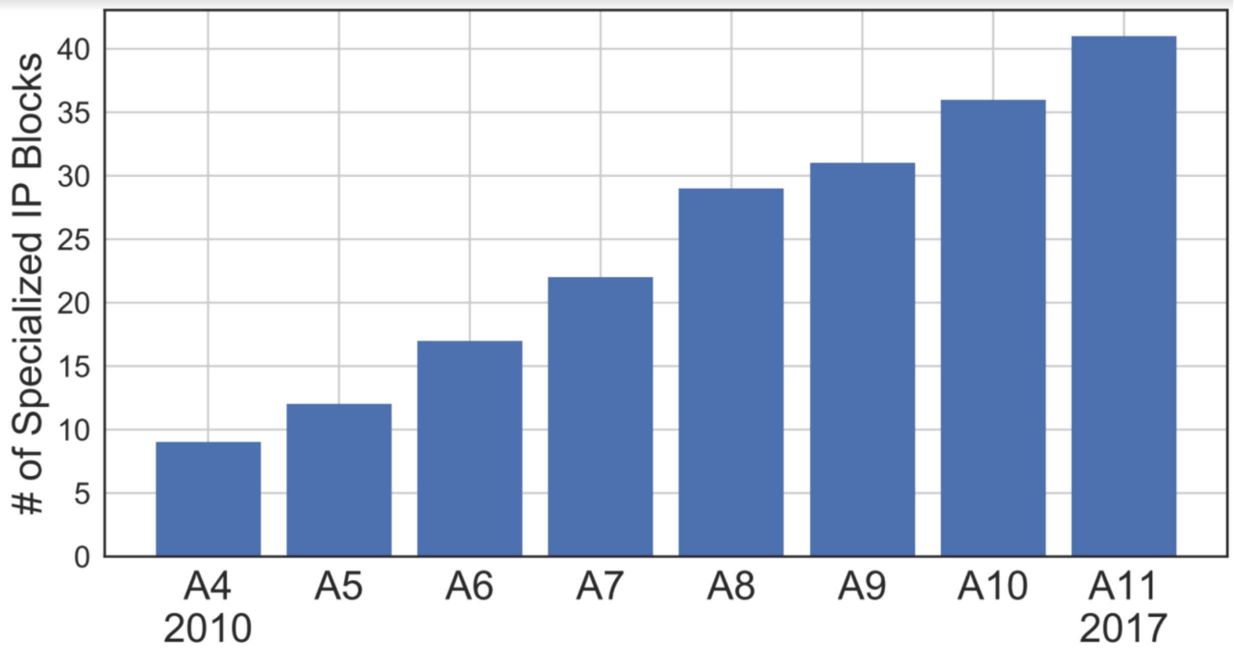


Xilinx ACAP

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

# Analysis of Apple A-* SoCs





http://vlsiarch.eecs.harvard.edu/accelerators/die-photo-analysis

# Memory Hierarchy is Specializing, Expanding, and Diversifying

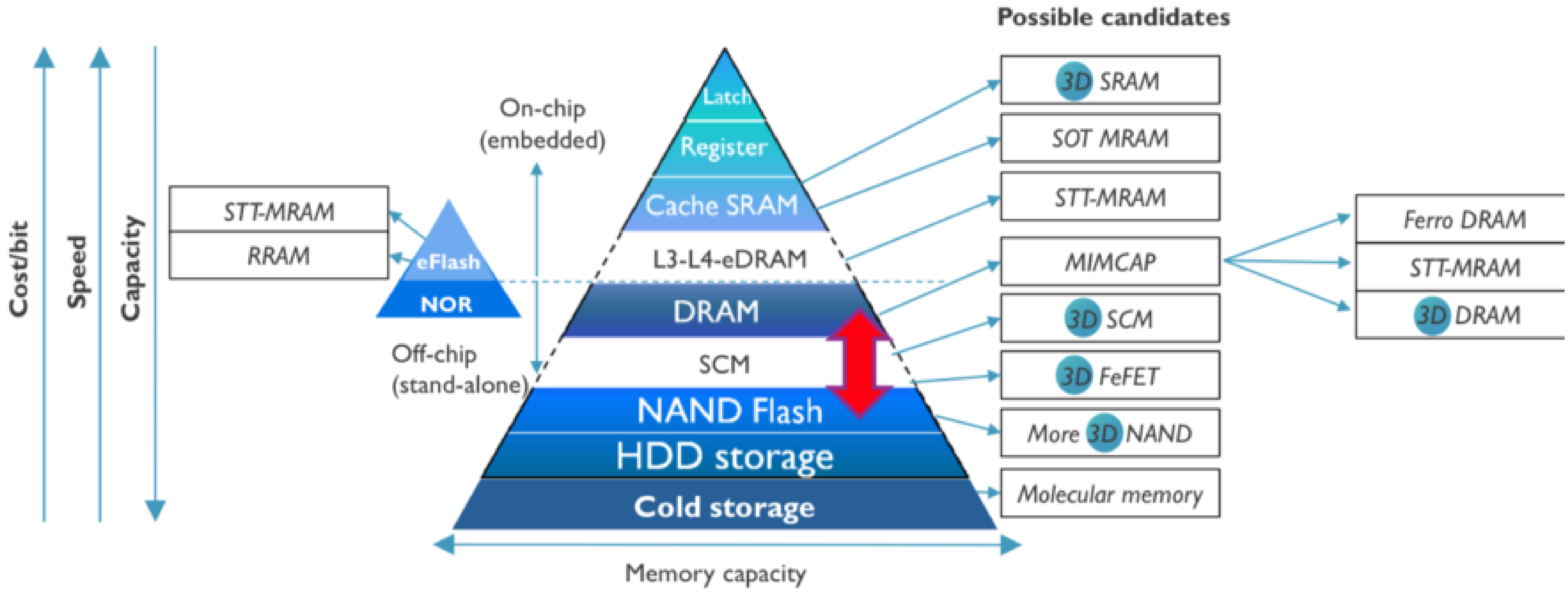
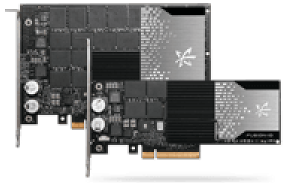
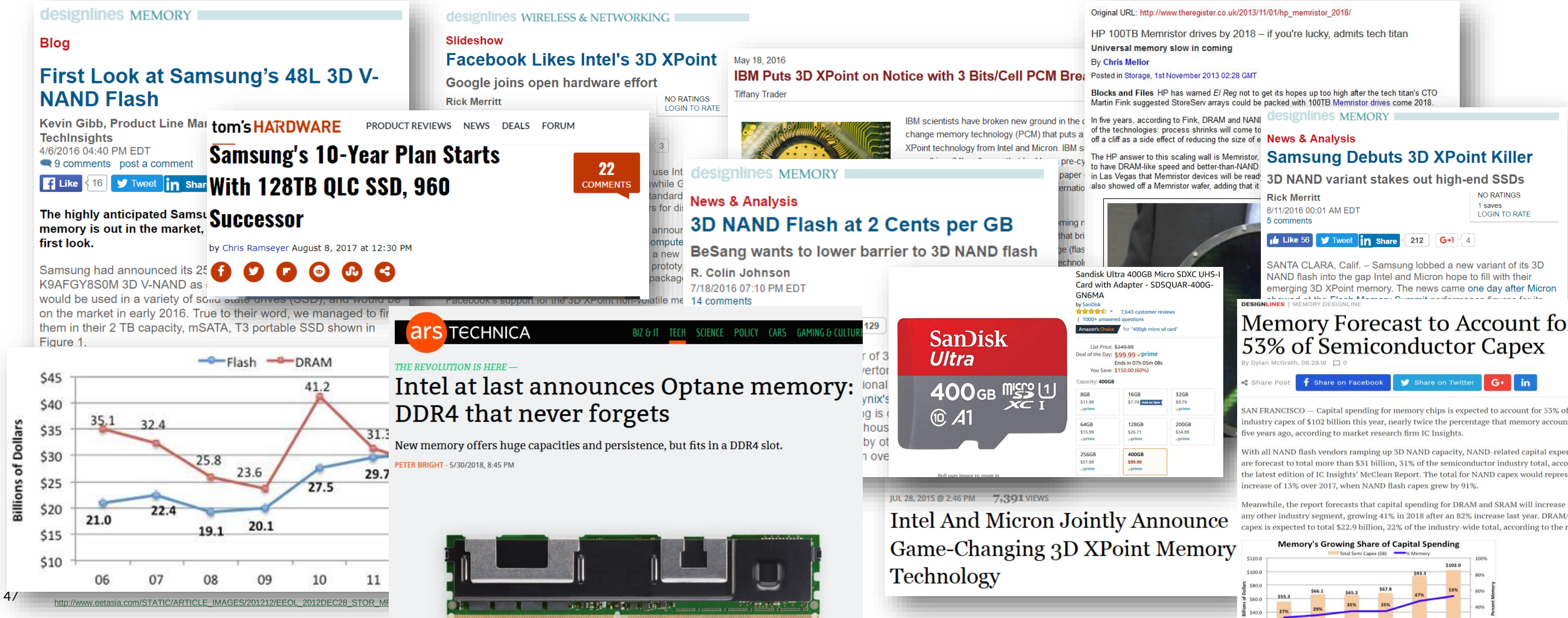


Image Source: IMEC

# 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 solid state drives (SSD), and would be 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.

http://www.eetasia.com/STATIC/ARTICLE_IMAGES/201212/EEOL_2012DEC28_STOR_MI

designlines WIRELESS & NETWORKING

Slideshow

## Facebook Likes Intel's 3D XPoint

Google joins open hardware effort

Rick Merritt

NO RATINGS
LOGIN TO RATE

3

use Int
while G
tandard
rs for di
annour
compute
a new
prototy
packag

Facebook's support for the 3D XPoint non-volatile me

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

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 c
change memory technology (PCM) that puts a
XPoint technology from Intel and Micron. IBM s
pre-cy
paper
ernatio
ming n
that bri
ge (flas
echnol

designlines MEMORY

News & Analysis

## 3D NAND Flash at 2 Cents per GB

BeSang wants to lower barrier to 3D NAND flash

R. Colin Johnson
7/18/2016 07:10 PM EDT
14 comments

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 read
also showed off a Memristor wafer, adding that it

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

SANTA CLARA, Calif. – Samsung lobbed a new variant of its 3D NAND flash into the gap Intel and Micron hope to fill with their emerging 3D XPoint memory. The news came one day after Micron

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

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

129

r of 3
verton
ional
ynix's
ng is
hous
by ot
n ove

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<br>$11.99<br>✓prime | 16GB<br>$7.79 Add-on Item | 32GB<br>$9.79<br>✓prime |
| 64GB<br>$15.99<br>✓prime | 128GB<br>$26.71<br>✓prime | 200GB<br>$34.99<br>✓prime |
| 256GB<br>$51.99<br>✓prime | 400GB<br>$99.99<br>✓prime | |

Roll over image to zoom in

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 Account fo 53% of Semiconductor Capex

By Dylan McGrath, 08.29.18 0

Share Post Share on Facebook Share on Twitter G+ in

SAN FRANCISCO — Capital spending for memory chips is expected to account for 53% of industry capex of $102 billion this year, nearly twice the percentage that memory accoun five years ago, according to market research firm IC Insights.

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

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

Memory's Growing Share of Capital Spending

# 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

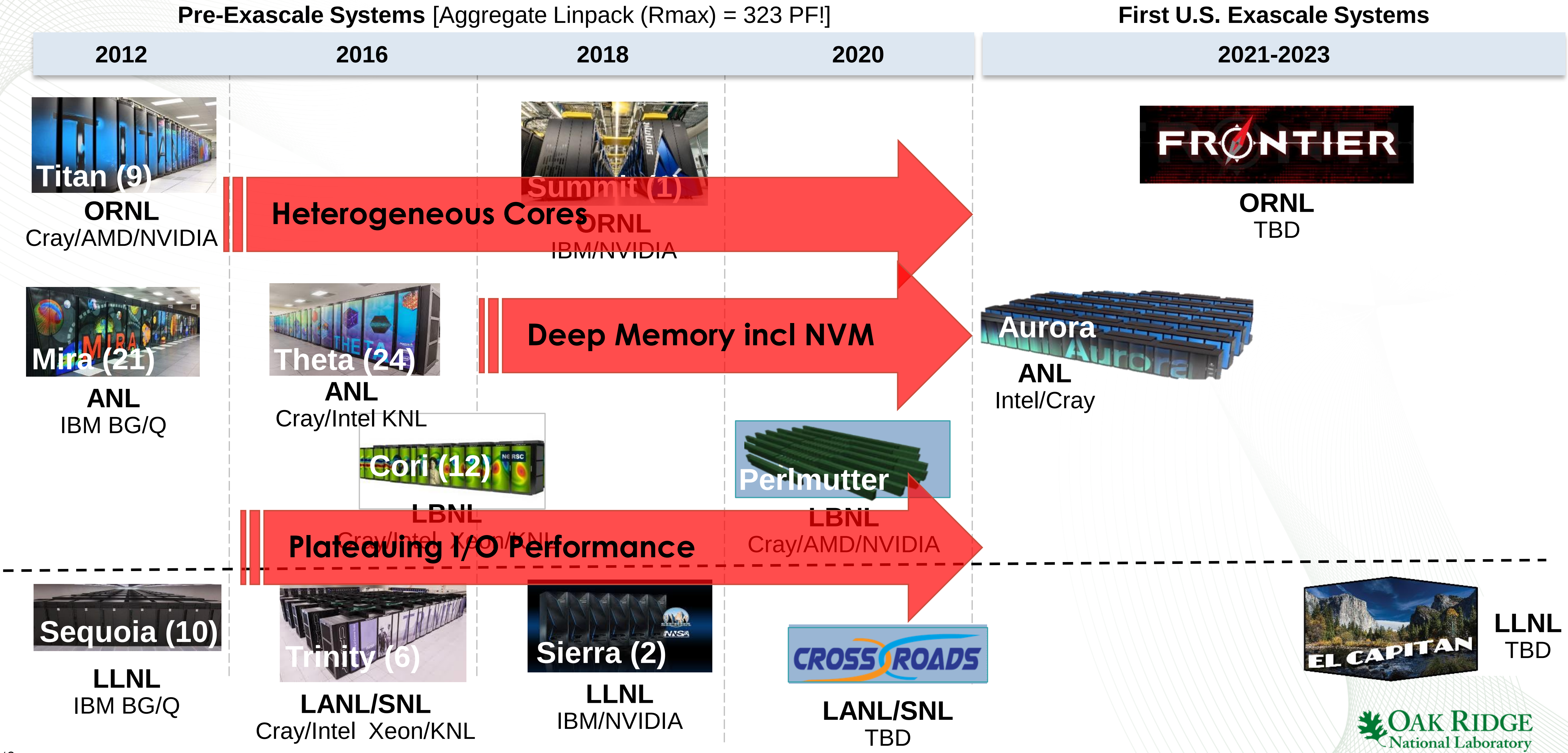
| Layer | Switch, 3D | NVM | Approximate | Neuro | Quantum |
|---|---|---|---|---|---|
| *Application* | 1 | 1 | 2 | 2 | 3 |
| *Algorithm* | 1 | 1 | 2 | 3 | 3 |
| *Language* | 1 | 2 | 2 | 3 | 3 |
| *API* | 1 | 2 | 2 | 3 | 3 |
| *Arch* | 1 | 2 | 2 | 3 | 3 |
| *ISA* | 1 | 2 | 2 | 3 | 3 |
| *Microarch* | 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

- Solutions need a co-design framework to evaluate and mature specific technologies

# 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 |
|---|---|---|---|---|
| Titan (9)<br>**ORNL**<br>Cray/AMD/NVIDIA | | Summit (1)<br>**ORNL**<br>IBM/NVIDIA | | FRONTIER<br>**ORNL**<br>TBD |
| Mira (21)<br>**ANL**<br>IBM BG/Q | Theta (24)<br>**ANL**<br>Cray/Intel KNL | | | Aurora<br>**ANL**<br>Intel/Cray |
| | Cori (12)<br>**LBNL**<br>Cray/Intel Xeon/KNL | | Perlmutter<br>**LBNL**<br>Cray/AMD/NVIDIA | |
| Sequoia (10)<br>**LLNL**<br>IBM BG/Q | Trinity (6)<br>**LANL/SNL**<br>Cray/Intel Xeon/KNL | Sierra (2)<br>**LLNL**<br>IBM/NVIDIA | CROSSROADS<br>**LANL/SNL**<br>TBD | EL CAPITAN<br>**LLNL**<br>TBD |

Heterogeneous Cores

Deep Memory incl NVM

Plateauing I/O Performance

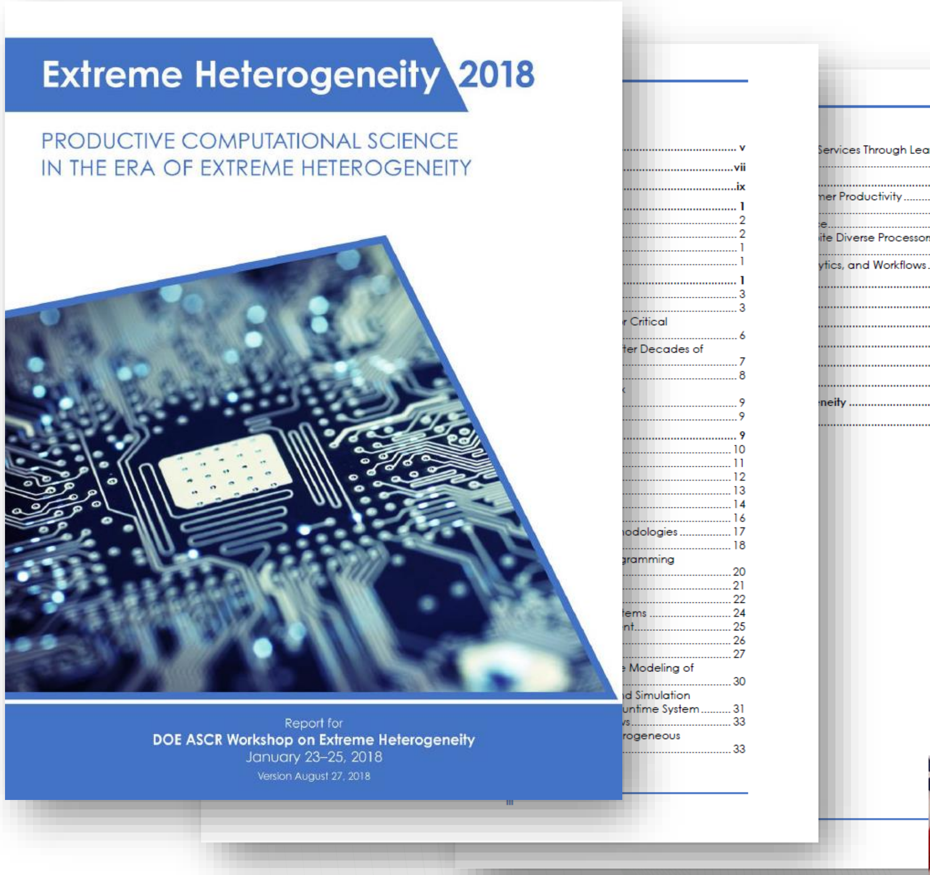
Oak Ridge National Laboratory

Jan 2018

# 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



Extreme Heterogeneity 2018

PRODUCTIVE COMPUTATIONAL SCIENCE IN THE ERA OF EXTREME HETEROGENEITY

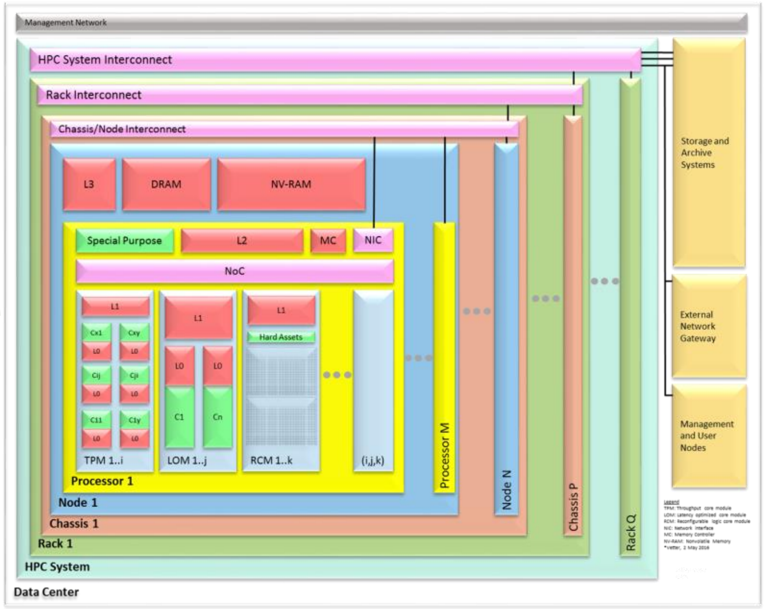
Report for
DOE ASCR Workshop on Extreme Heterogeneity
January 23–25, 2018
Version August 27, 2018

https://orau.gov/exheterogeneity2018/

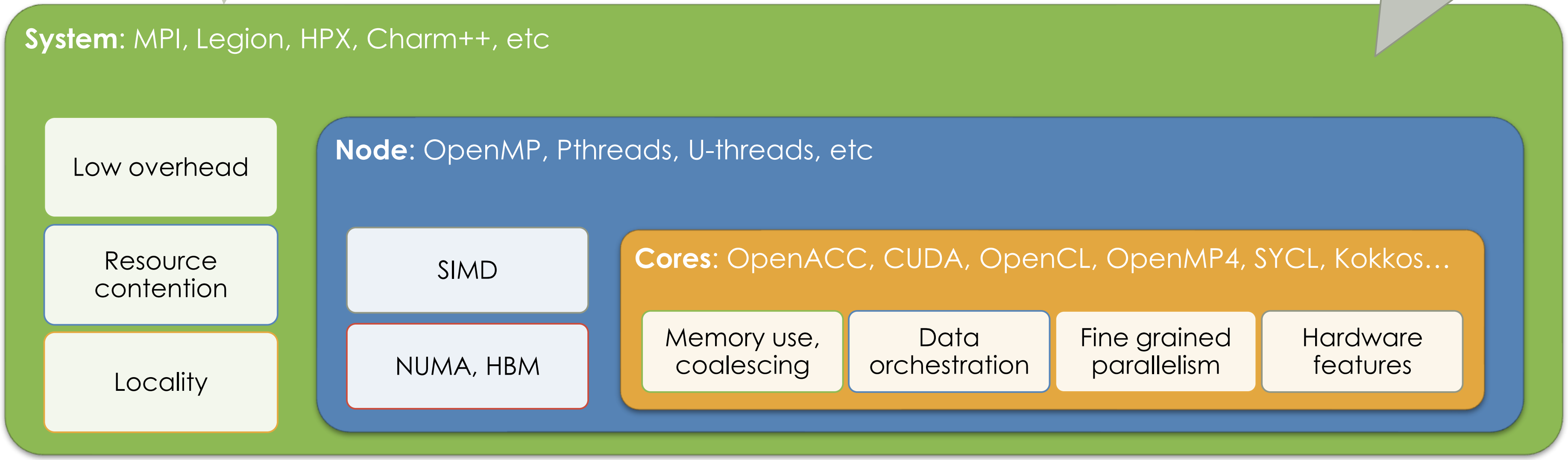
https://doi.org/10.2172/1473756

# Programming Heterogeneous Systems

# Complex Architectures Yields Complex Programming Models



- This approach is not scalable, affordable, robust, elegant, etc.
- Not performance portable across different architectures



**System**: MPI, Legion, HPX, Charm++, etc

Low overhead

Resource contention

Locality

**Node**: OpenMP, Pthreads, U-threads, etc

SIMD

NUMA, HBM

**Cores**: OpenACC, CUDA, OpenCL, OpenMP4, SYCL, Kokkos…

Memory use, coalescing

Data orchestration

Fine grained parallelism

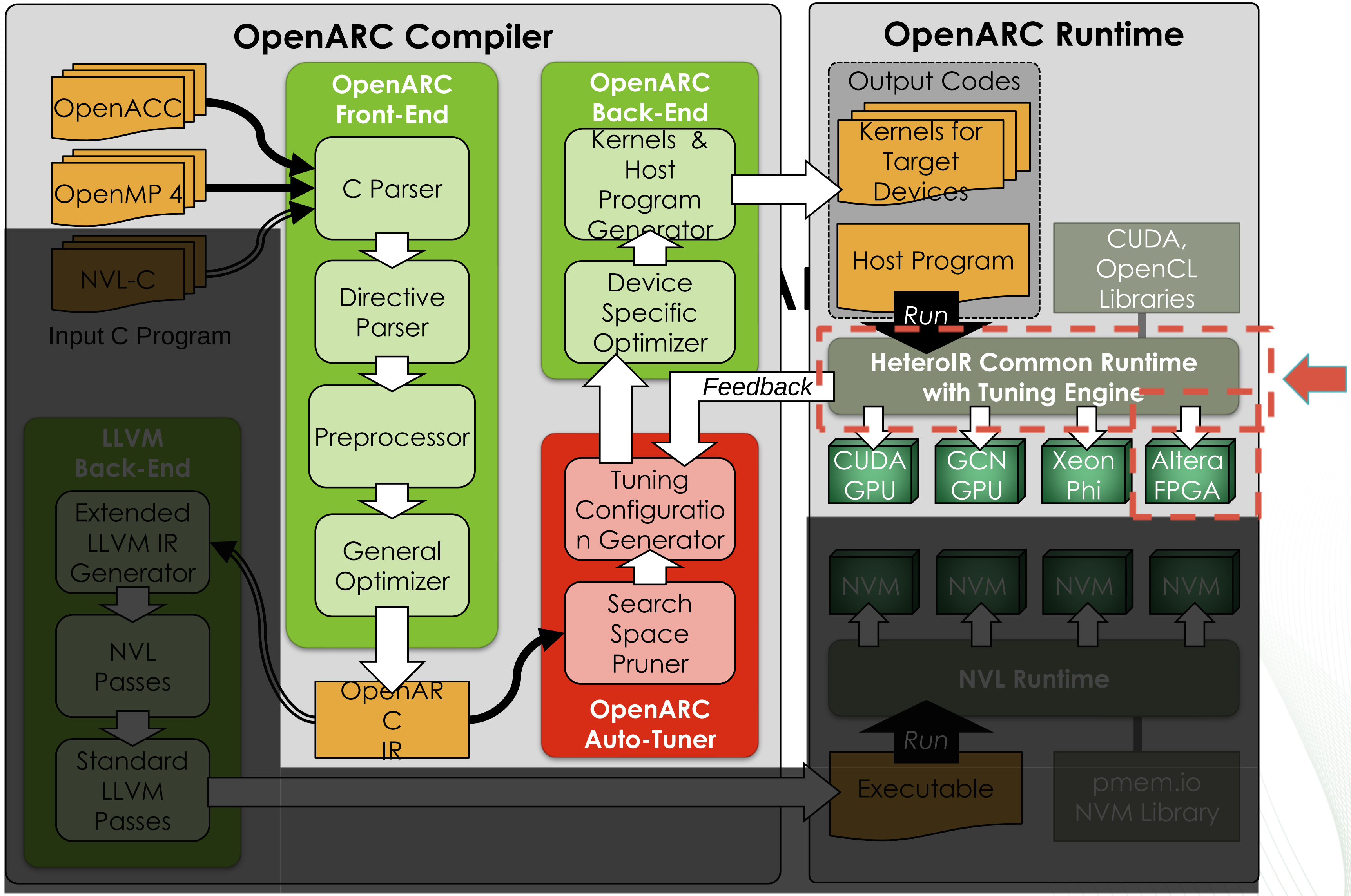
Hardware features

# Directive-based Solutions for FPGA Computing

# FPGAs| Approach

- Design and implement an OpenACC-to-FPGA translation framework, which is the first work to use a standard and portable directive-based, high-level programming system for FPGAs.
- Propose FPGA-specific optimizations and novel pragma extensions to improve performance.
- Evaluate the functional and performance portability of the framework across diverse architectures (Altera FPGA, NVIDIA GPU, AMD GPU, and Intel Xeon Phi).

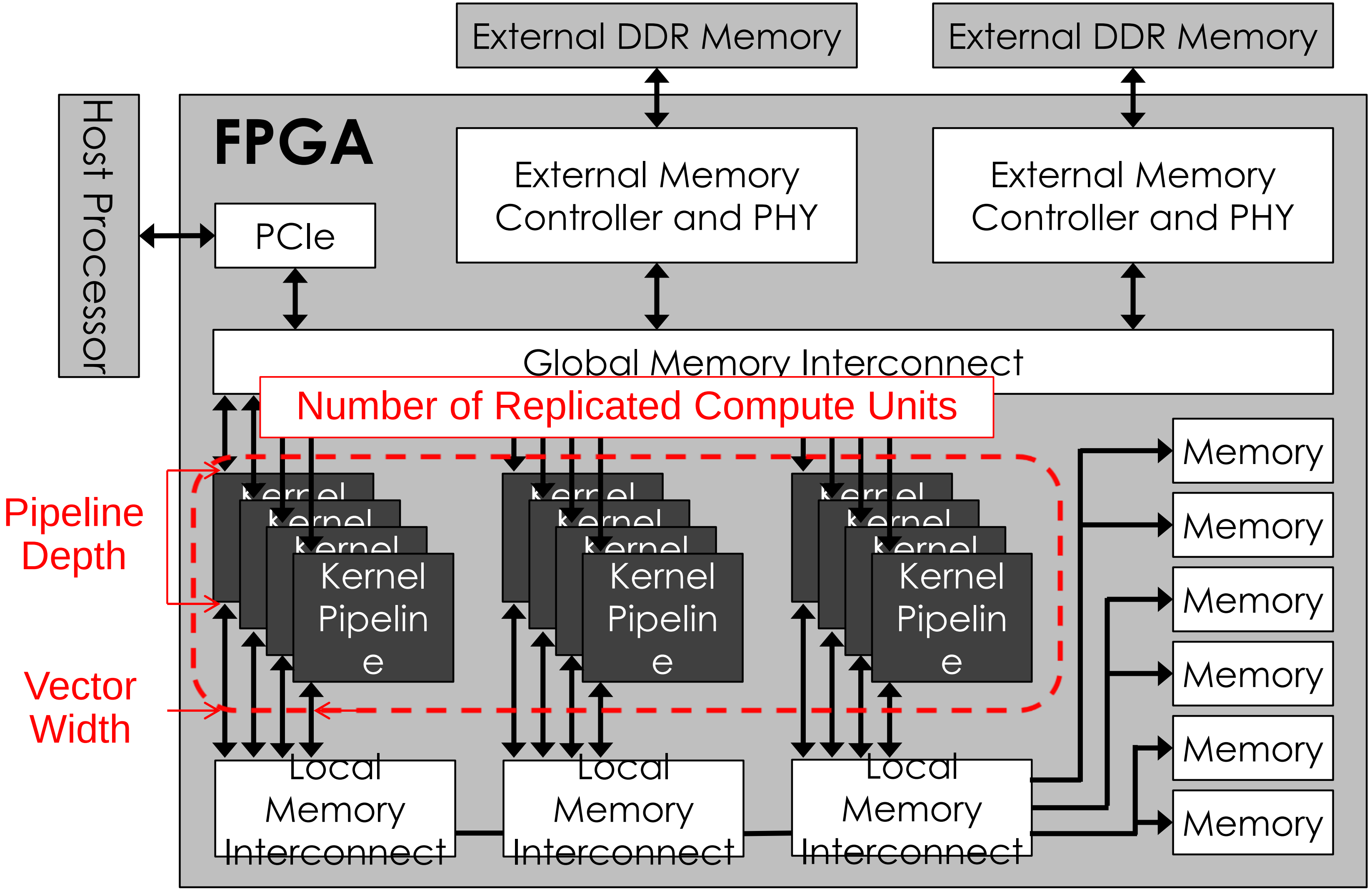
S. Lee, J. Kim, and J.S. Vetter, "OpenACC to FPGA: A Framework for Directive-based High-Performance Reconfigurable Computing," Proc. IEEE International Parallel & Distributed Processing Symposium (IPDPS), 2016, 10.1109/IPDPS.2016.28.


OpenARC Compiler
OpenACC
OpenMP 4
NVL-C
Input C Program
OpenARC Front-End
C Parser
Directive Parser
Preprocessor
General Optimizer
OpenARC C IR
OpenARC Back-End
Kernels & Host Program Generator
Device Specific Optimizer
Feedback
Tuning Configuration Generator
Search Space Pruner
OpenARC Auto-Tuner
LLVM Back-End
Extended LLVM IR Generator
NVL Passes
Standard LLVM Passes
OpenARC Runtime
Output Codes
Kernels for Target Devices
Host Program
Run
CUDA, OpenCL Libraries
HeteroIR Common Runtime with Tuning Engine
CUDA GPU
GCN GPU
Xeon Phi
Altera FPGA
NVM
NVM
NVM
NVM
NVL Runtime
Run
Executable
pmem.io NVM Library

# Baseline Translation of OpenACC-to-FPGA

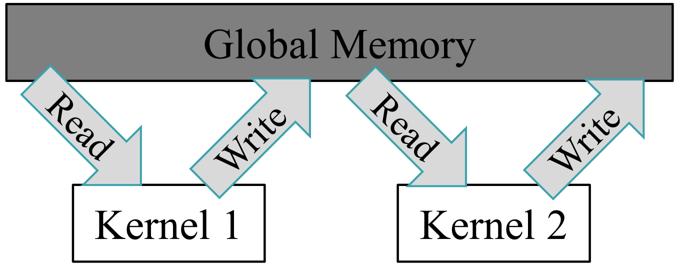
- Use OpenCL as the output model and the Altera Offline Compiler (AOC) as its backend compiler.
- Translates the input OpenACC program into a host code containing HeteroIR constructs and device-specific kernel codes.
  - Use the same HeteroIR runtime system of the existing OpenCL backends, except for the device initialization.
  - Reuse most of compiler passes for kernel generation.

# FPGA OpenCL Architecture

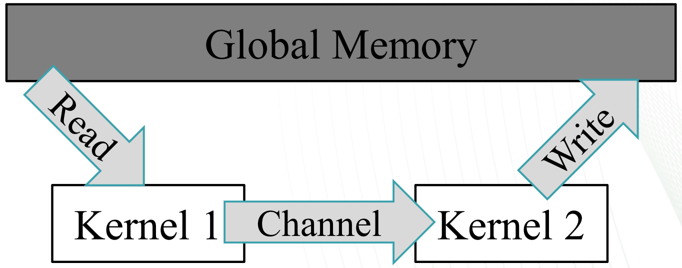

External DDR Memory
External DDR Memory
Host Processor
FPGA
PCIe
External Memory Controller and PHY
External Memory Controller and PHY
Global Memory Interconnect
Number of Replicated Compute Units
Pipeline Depth
Vector Width
Kernel Pipeline
Kernel Pipeline
Kernel Pipeline
Local Memory Interconnect
Local Memory Interconnect
Local Memory Interconnect
Memory
Memory
Memory
Memory
Memory
Memory

# Kernel-Pipelining Transformation Optimization

- Kernel execution model in OpenACC
  - Device kernels can communicate with each other only through the device global memory.
  - Synchronizations between kernels are at the granularity of a kernel execution.
- Altera OpenCL channels
  - Allows passing data between kernels and synchronizing kernels with high efficiency and low latency



Kernel communications through global memory in OpenACC

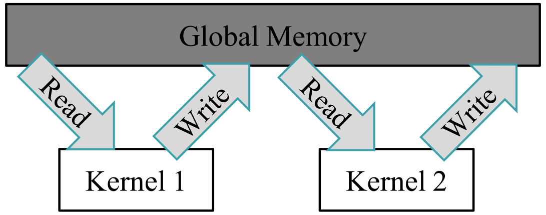


Kernel communications with Altera channels

# Kernel-Pipelining Transformation Optimization (2)

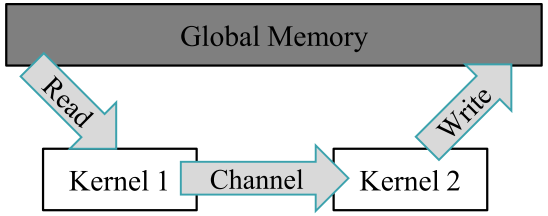
(a) Input OpenACC code

```
#pragma acc data copyin (a) create (b) copyout (c)
{
    #pragma acc kernels loop gang worker present (a, b)
    for(i=0; i<N; i++) { b[i] = a[i]*a[i]; }
    #pragma acc kernels loop gang worker present (b, c)
    for(i=0; i<N; i++) {c[i] = b[i]; }
}
```



(b) Altera OpenCL code with channels

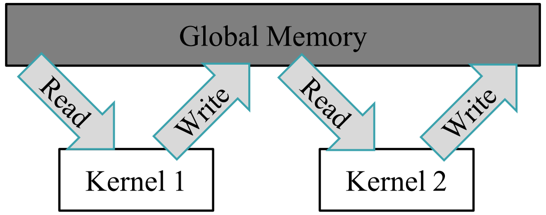
```
channel float pipe_b;
__kernel void kernel1(__global float* a) {
    int i = get_global_id(0);
    write_channel_altera(pipe_b, a[i]*a[i]);
}
__kernel void kernel2(__global float* c) {
    int i = get_global_id(0);
    c[i] = read_channel_altera(pipe_b);
}
```



# Kernel-Pipelining Transformation Optimization (3)

(a) Input OpenACC code

```
#pragma acc data copyin (a) create (b) copyout (c)
{
    #pragma acc kernels loop gang worker present (a, b)
    for(i=0; i<N; i++) { b[i] = a[i]*a[i]; }
    #pragma acc kernels loop gang worker present (b, c)
    for(i=0; i<N; i++) {c[i] = b[i]; }
}
```

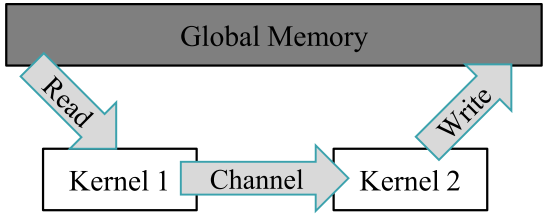


Kernel-pipelining transformation

Valid under specific conditions

(c) Modified OpenACC code for kernel-pipelining

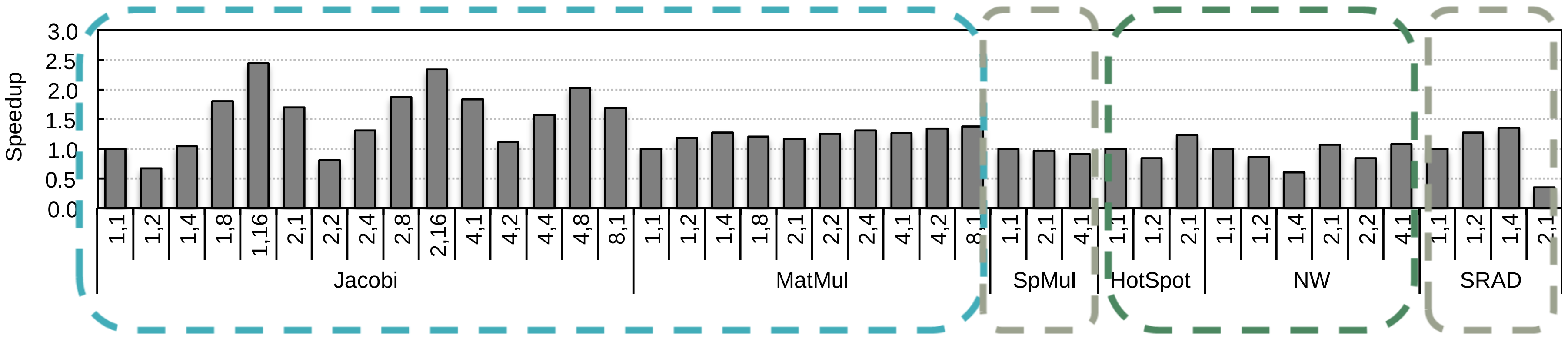
```
#pragma acc data copyin (a) pipe (b) copyout (c)
{
    #pragma acc kernels loop gang worker pipeout (b) present (a)
    For(i=0; i<N; i++) { b[i] = a[i]*a[i]; }
    #pragma acc kernels loop gang worker pipein (b) present (c)
    For(i=0; i<N; i++) {c[i] = b[i];}
}
```



# Speedup over CU, SIMD (1,1)

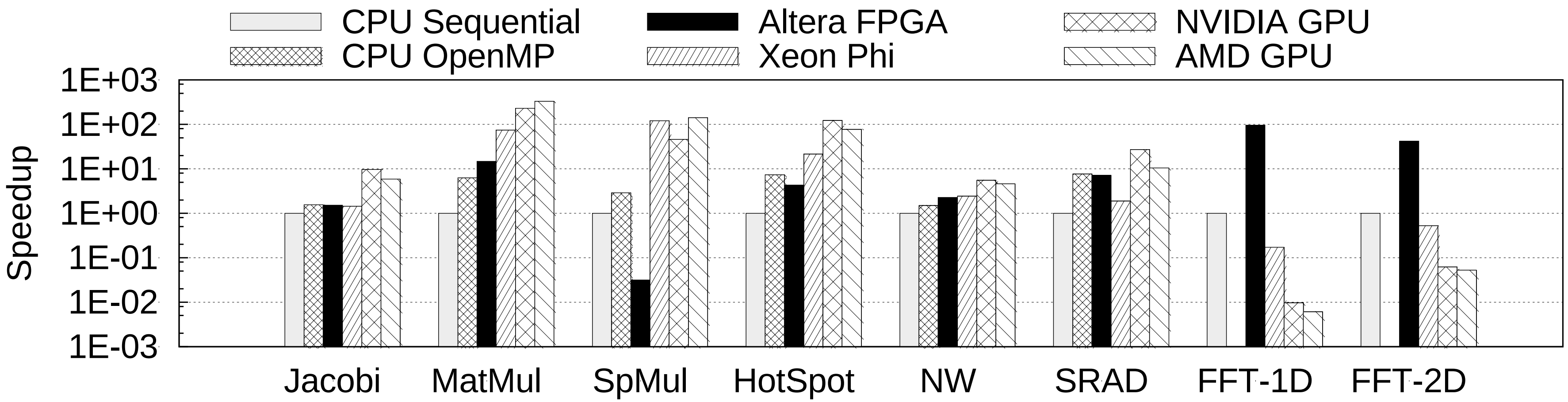
Jacobi and MatMul show better performance with increase in CU and SIMD, thanks to regular memory accesses.

SpMul and SRAD perform worse with multiple CUs, mainly due to memory contention.



Performance of HotSpot and NW increases with multiple CUs, but decreases with vectorization.

# Overall Performance



FPGAs prefer applications with deep execution pipelines (e.g., FFT-1D and FFT-2D), performing much higher than other accelerators.

For traditional HPC applications with abundant parallel floating-point operations, it seems to be difficult for FPGAs to beat the performance of other accelerators, even though FPGAs can be much more power-efficient.

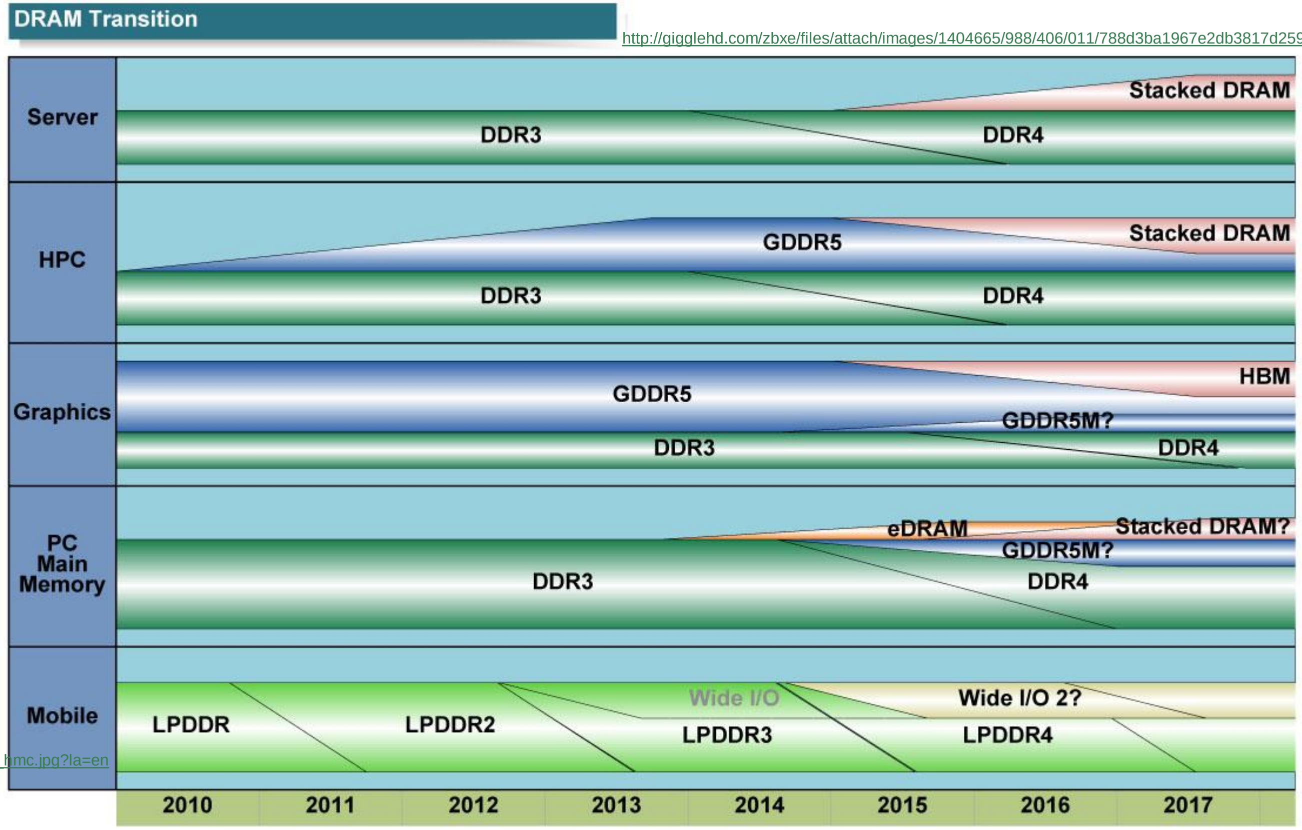
- Tested FPGA does not contain dedicated, embedded floating-point cores, while others have fully-optimized floating-point computation units.

Current and upcoming high-end FPGAs are equipped with hardened floating-point operators, whose performance will be comparable to other accelerators, while remaining power-efficient.

# Emerging Memory Systems

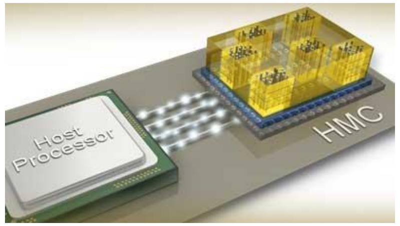
# Memory Systems Started Diversifying Several Years Ago

- 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

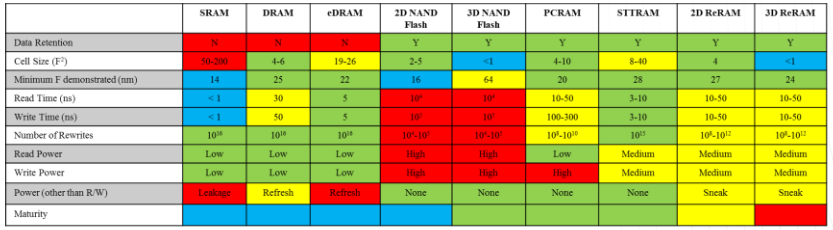


http://gigglehd.com/zbxe/files/attach/images/1404665/988/406/011/788d3ba1967e2db3817d259d2e

Copyright (c) 2014 Hiroshige Goto All rights reserved.



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



| | 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.*

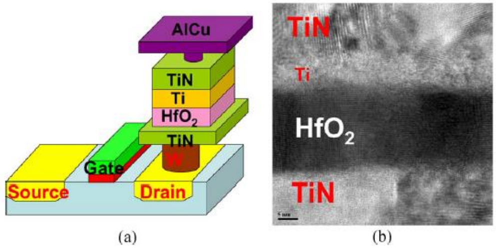


**Fig. 4.** *(a) A typical 1T1R structure of RRAM with $HfO_x$; (b) HR-TEM image of the TiN/Ti/$HfO_x$/TiN stacked layer; the thickness of the $HfO_2$ is 20 nm.*

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

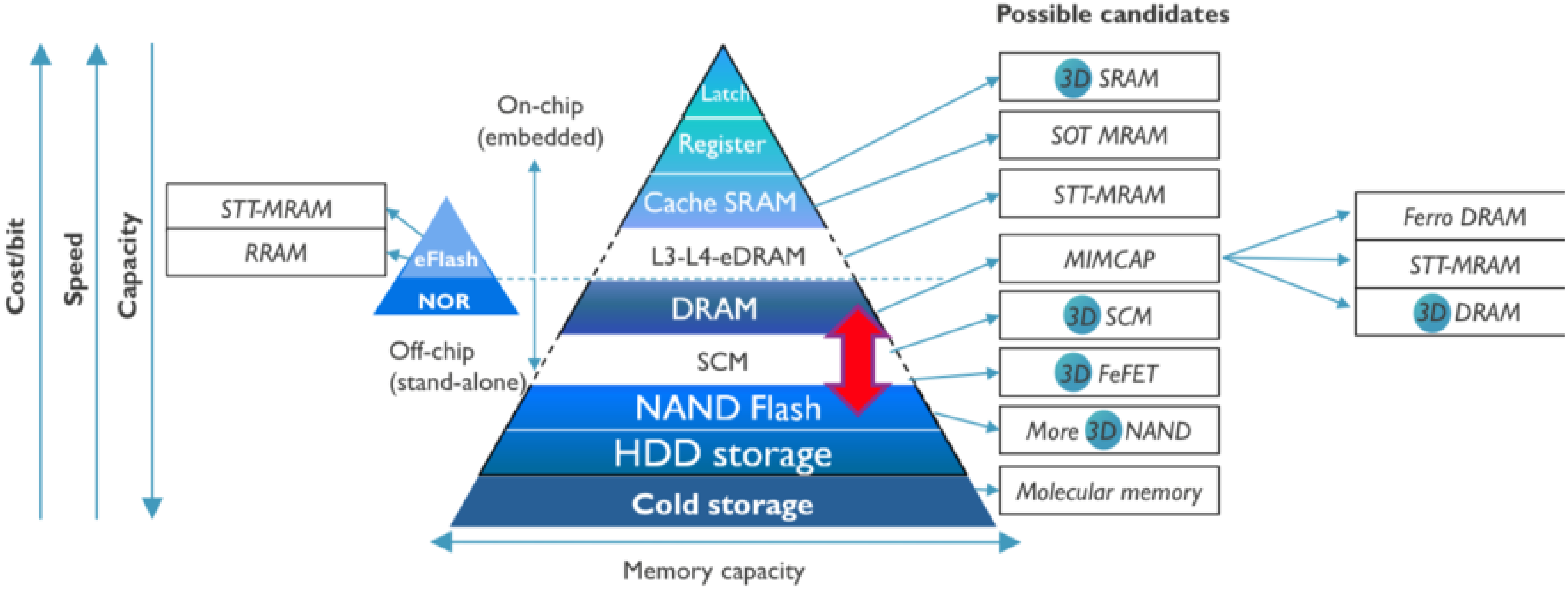
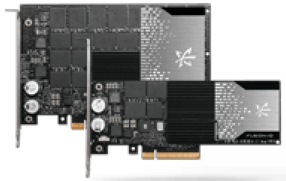
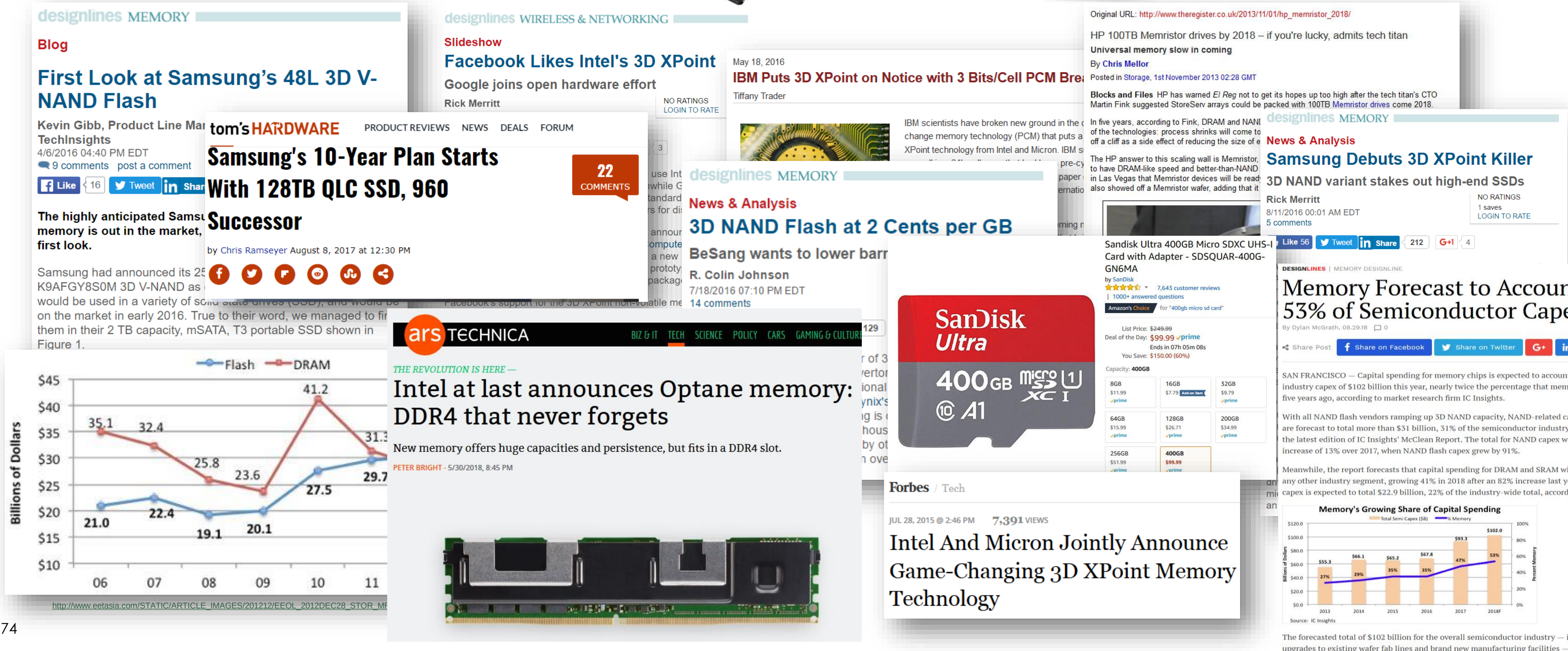


Image Source: IMEC

# 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 so
on the market in early 2016. True to their word, we managed to fir
them in their 2 TB capacity, mSATA, T3 portable SSD shown in Figure 1.

http://www.eetasia.com/STATIC/ARTICLE_IMAGES/201212/EEOL_2012DEC28_STOR_MF

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 WIRELESS & NETWORKING

Slideshow

Facebook Likes Intel's 3D XPoint

Google joins open hardware effort

Rick Merritt

NO RATINGS
LOGIN TO RATE

use Int
while G
tandard
rs for di
annour
ompute
a new
prototy
packag
Facebook's support for the 3D XPoint non-volatile me

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
pre-cy
paper
ernatio
ming

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

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 NAN
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 read
also showed off a Memristor wafer, adding that it

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

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

*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

129

r of 3
ertor
ional
ynix's
ng is
hous
by ot
n ove

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<br>$11.99<br>✓prime | 16GB<br>$7.79 Add-on Item | 32GB<br>$9.79<br>✓prime |
| 64GB<br>$15.99<br>✓prime | 128GB<br>$26.71<br>✓prime | 200GB<br>$34.99<br>✓prime |
| 256GB<br>$51.99<br>✓prime | 400GB<br>$99.99<br>✓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 Accour 53% of Semiconductor Cape

By Dylan McGrath, 08.29.18 0

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 ca
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

Memory's Growing Share of Capital Spending

Source: IC Insights

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

# Many Memory Architecture Options under Consideration…

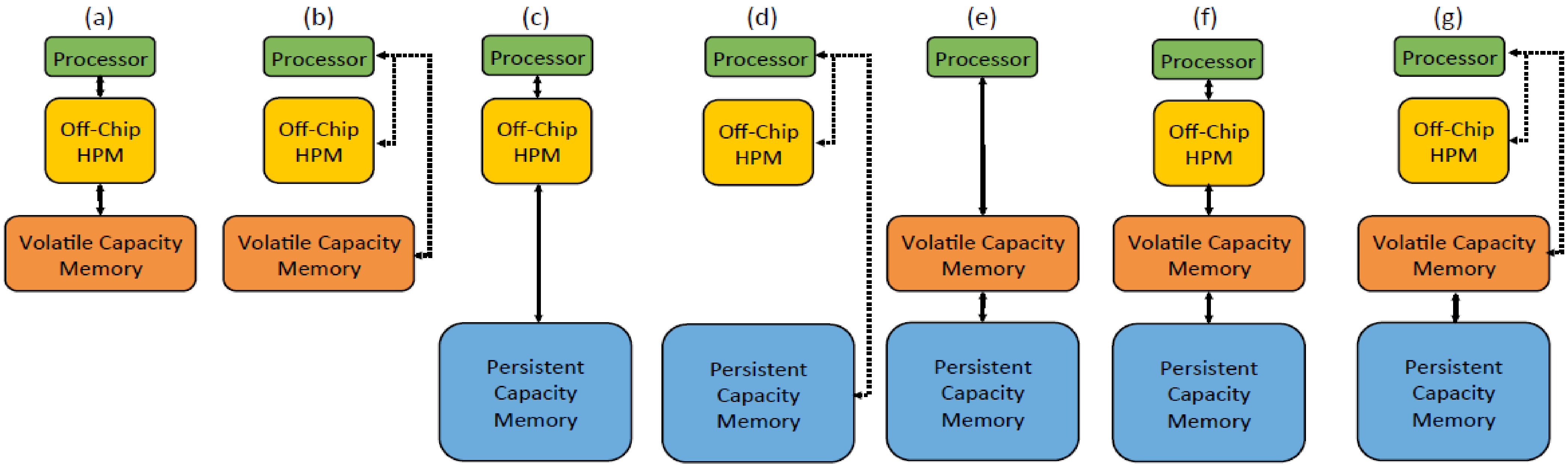


TABLE I: Comparison of four tiers of recent memory technologies [9], [11], [17], [18], [22]–[25], [28], [30], [35], [39], [40], [47]–[49].

| | 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 |

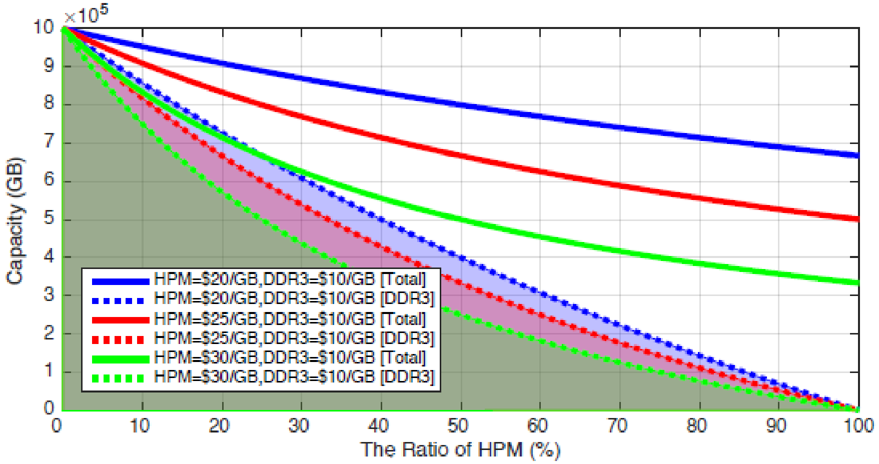


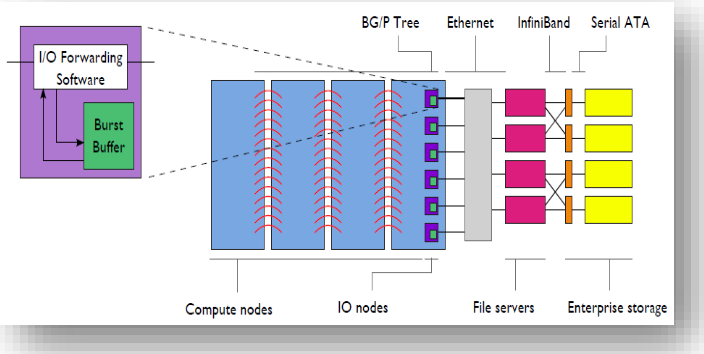
Fig. 1: Possible configurations of a memory system using DDR3 and HPM of different costs under a fixed budget.

RIDGE Laboratory

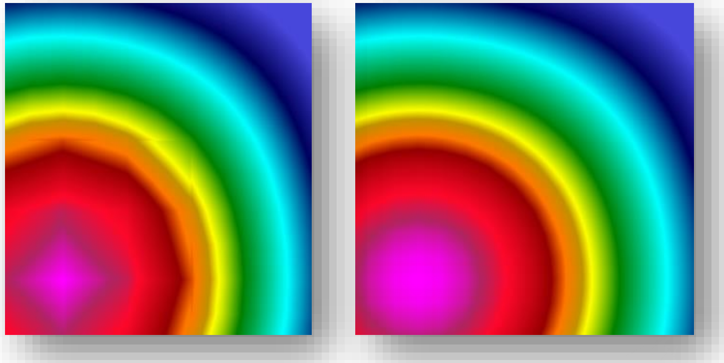
# Programming NVM Systems Portably

# 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

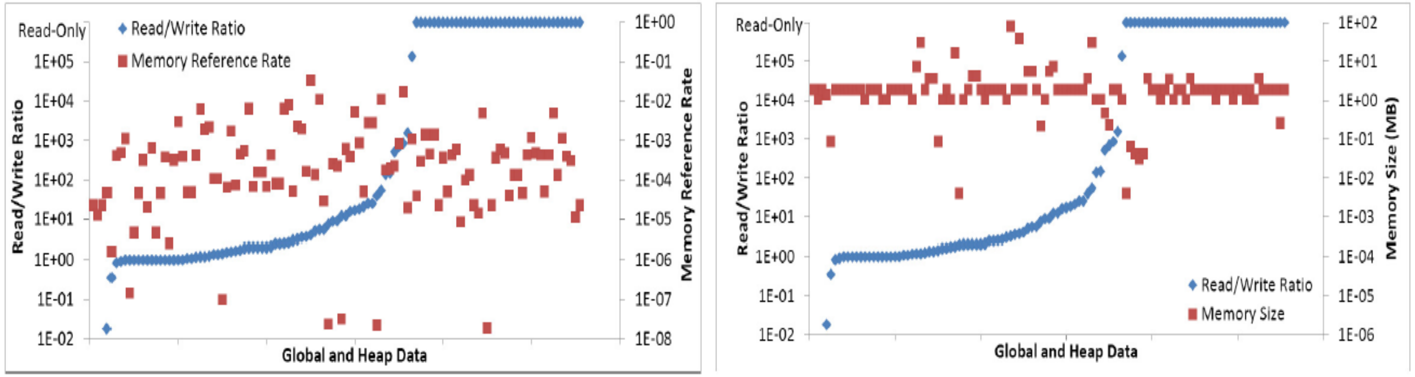


Figure 3: Read/write ratios, memory reference rates and memory object sizes for memory objects in Nek5000

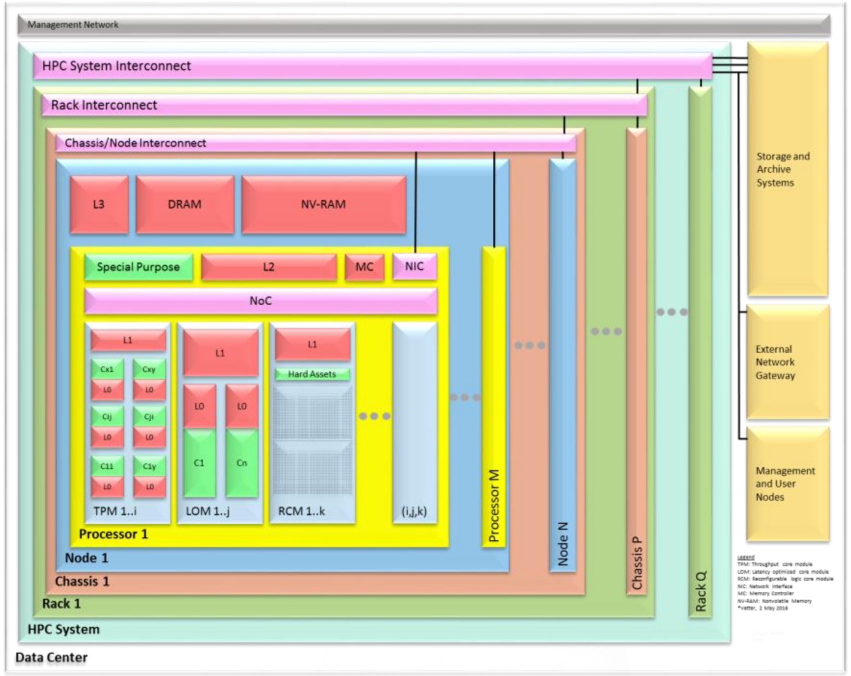
Empirical results show many reasons…

- 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…

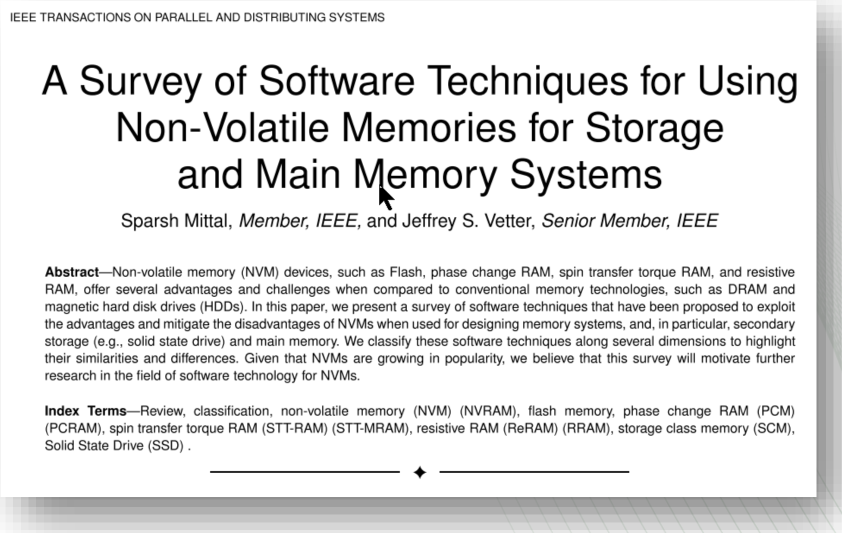
J.S. Vetter and S. Mittal, “Opportunities for Nonvolatile Memory Systems in Extreme-Scale High-Performance Computing,” *Computing in Science & Engineering, 17(2):73-82, 2015.*

# 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) .

# 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);
}
```

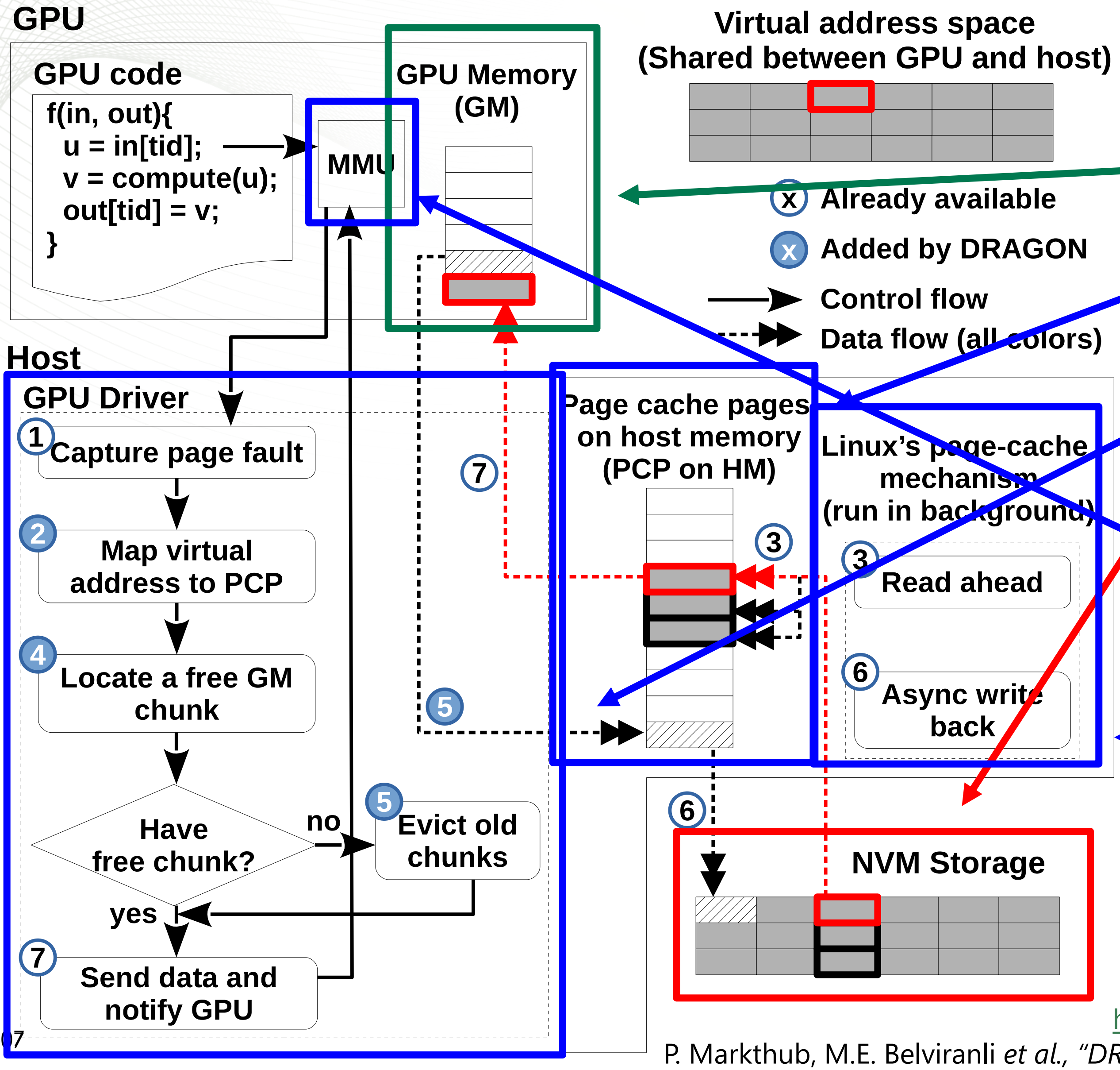
## DRAGON

```
// 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**

# 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*

# Results with Caffe

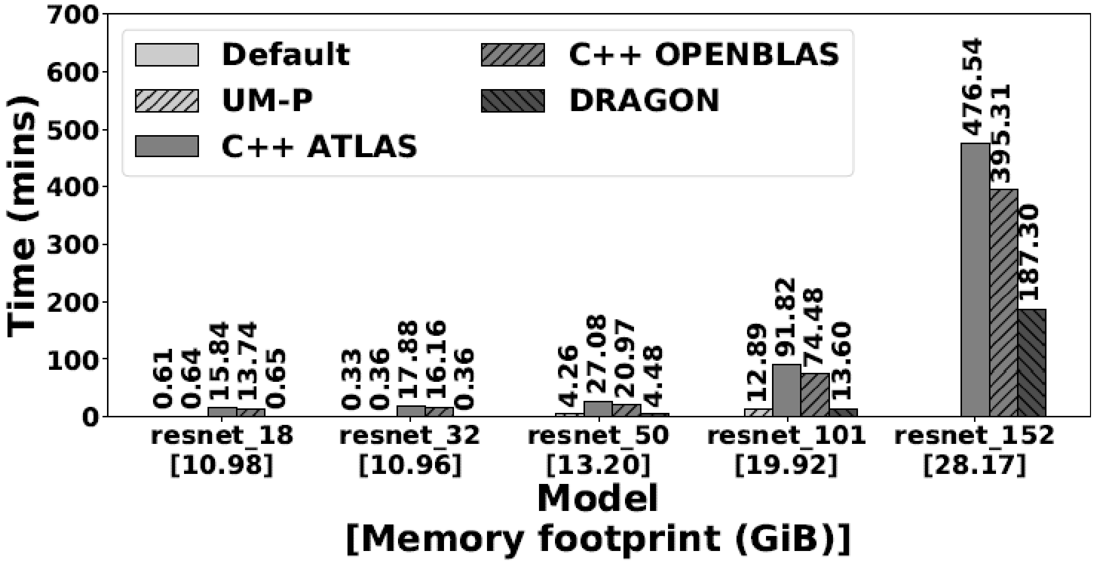


Figure 6: Comparison of ResNet execution times on Caffe.

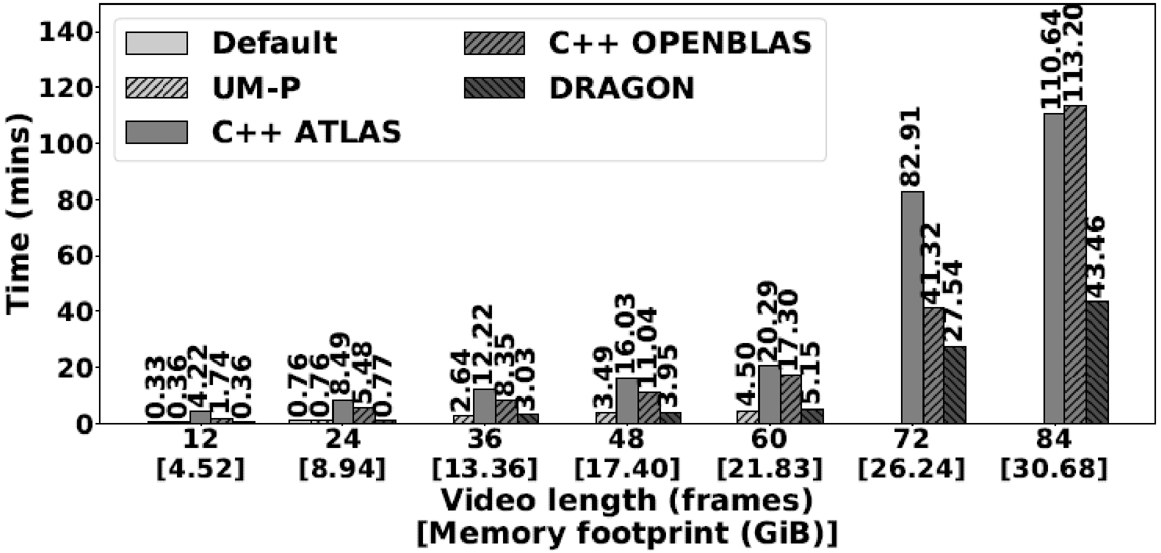


Figure 7: Comparison of C3D the execution times on Caffe.

- 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

OAK RIDGE National Laboratory

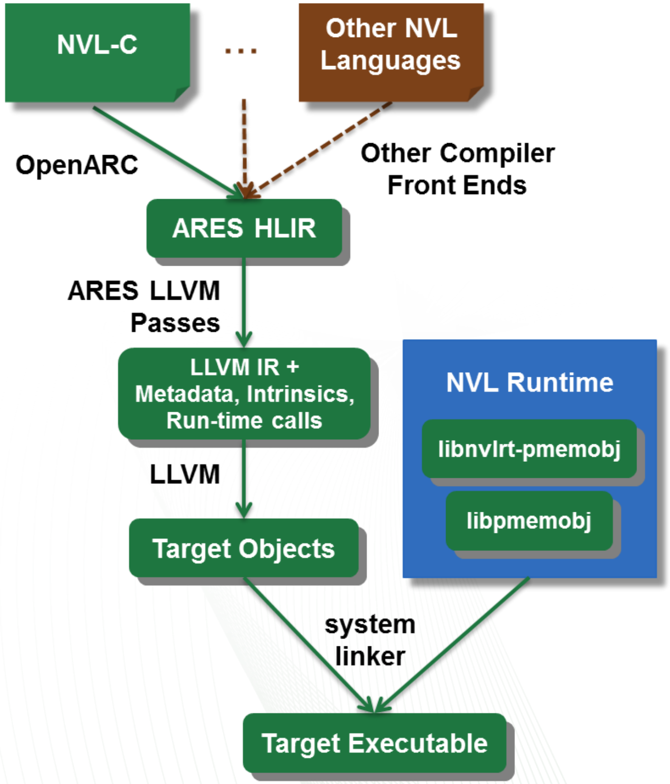
# NVL-C: Portable Programming for NVMM

- 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*

# Design Goals: Familiar programming interface

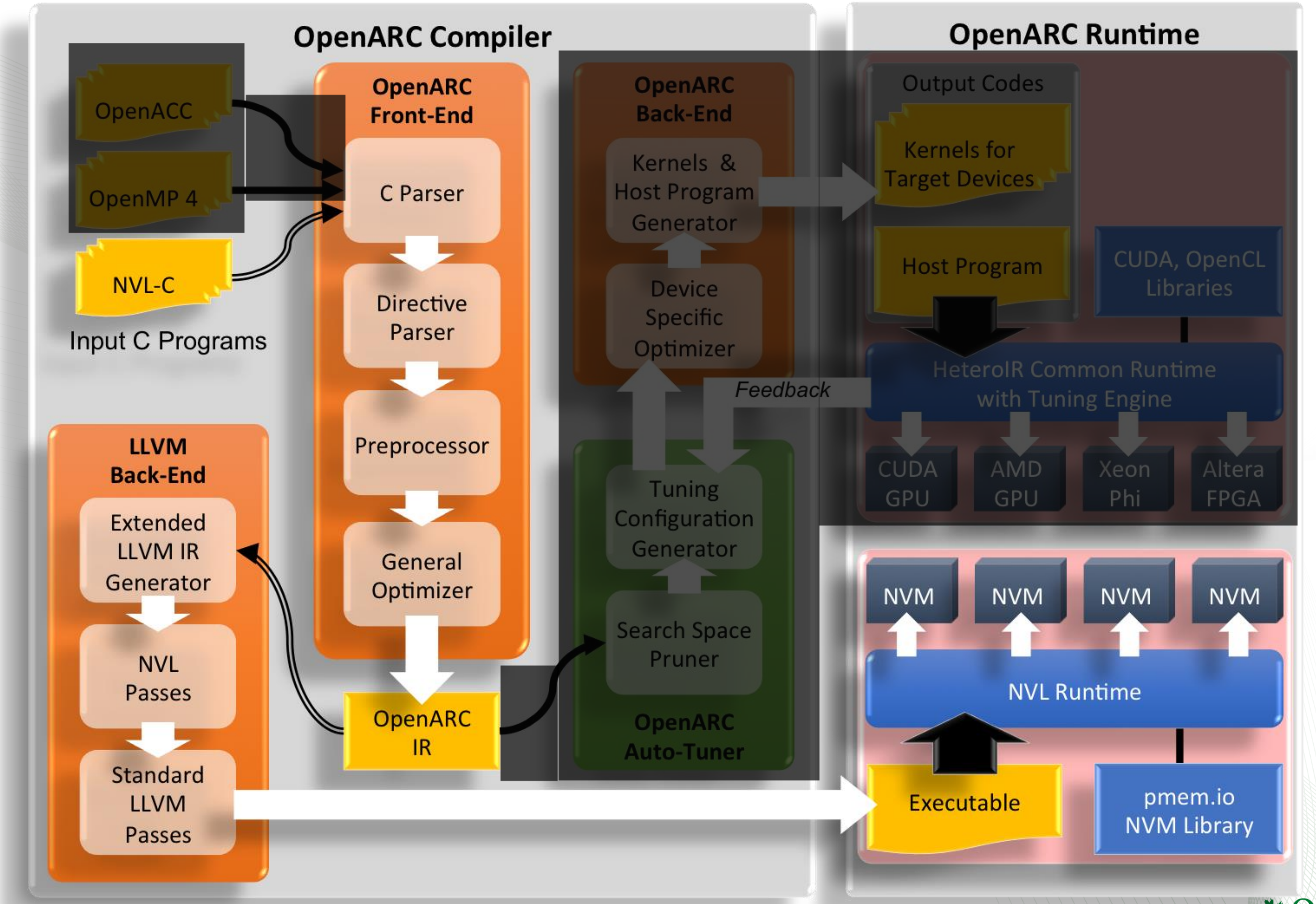
```
#include <nvl.h>
struct list {
  int value;
  nvl struct list *next;
};
void add(int k, nvl struct list *after) {
  nvl struct list *node
    = nvl_alloc_nv(heap, 1, struct list);
  node->value = k;
  node->next  = after->next;
  after->next = node;
}
```

- Small set of C language extensions:
  - Header file
  - Type qualifiers
  - Library API
  - Pragmas
- Existing memory interfaces remain:
  - NVL-C is a superset of C
  - Unqualified types as specified by C
  - Local/global variables stored in volatile memory (DRAM or registers)
  - Use existing C standard libraries for HDD

OAK RIDGE
National Laboratory

# Design Goals: Avoiding persistent data corruption

- New categories of pointer bugs:
  - Caused by multiple memory types:
    - E.g., pointer from NVM to volatile memory will become dangling pointer
  - Prevented at compile time or run time
- Automatic reference counting:
  - No need to manually free
  - Avoids leaks and dangling pointers
- Transactions:
  - Avoids persistent data corruption across software and hardware failures
- High performance:
  - Performance penalty from memory management, pointer safety, and transactions
  - Compiler-based optimizations
  - Programmer-specified hints



http://ft.ornl.gov/research/openarc

# Programming Model: NVM Pointers

```
#include <nvl.h>
struct list {
  int value;
  nvl struct list *next;
};
void add(int k, nvl struct list *after) {
  struct list *node
    = malloc(sizeof(struct list));
  node->value = k;
  node->next  = after->next;
  after->next = node;
}
```

*compile-time error*
*explicit cast won't help*

- `nvl` type qualifier:
  - Indicates NVM storage
  - On target type, declares NVM pointer
  - No NVM-stored local or global variable
- Stricter type safety for NVM pointers:
  - Does not affect other C types
  - Avoids persistent data corruption
  - Facilitates compiler analysis
  - Needed for automatic reference counting
  - E.g., pointer conversions involving NVM pointers are strictly prohibited

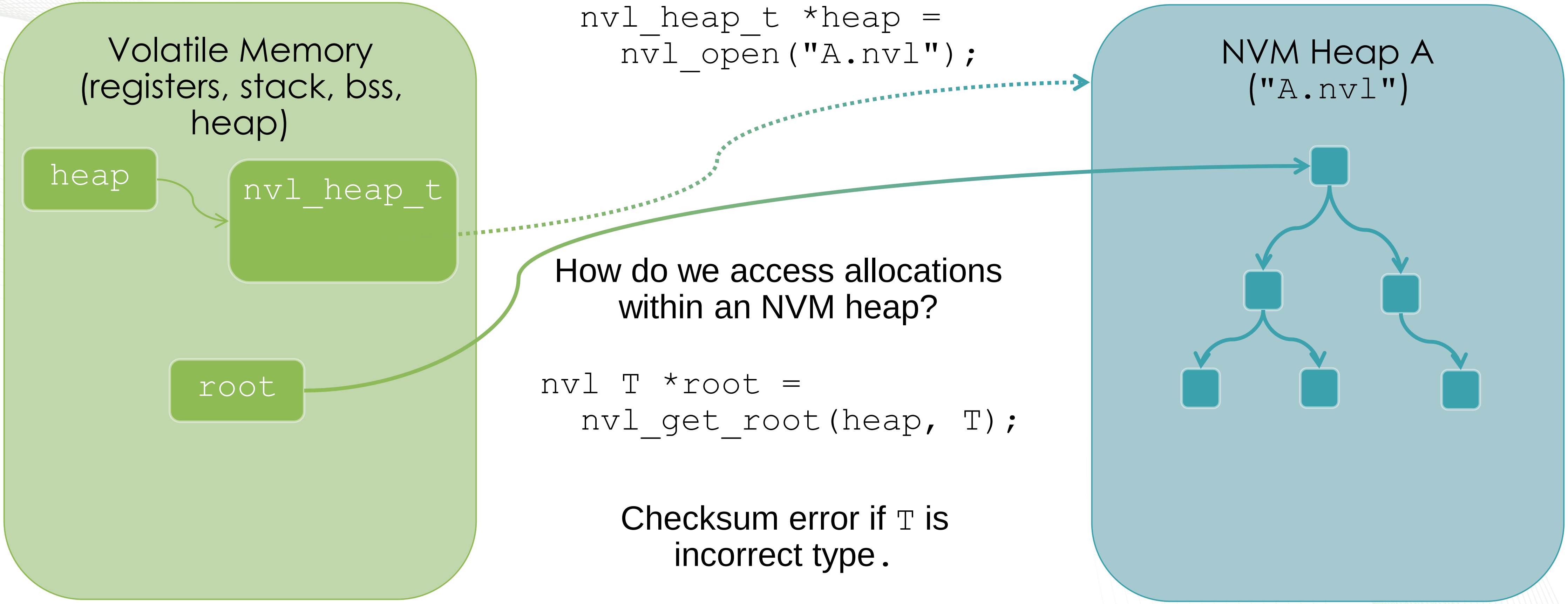
# Programming Model: NVM memory management

- Hybrid of traditional HDD and DRAM programming interfaces
- NVM storage organized into *NVM heaps* identified by file names
- NVM heaps can be managed using normal file system commands
- Within an NVM heap, memory always allocated dynamically

| NVM | HDD analogue |
|---|---|
| nvl_heap_t | FILE |
| nvl_open | fopen |
| nvl_close | fclose |
| mv, rm, ls, *etc.* | mv, rm, ls, *etc.* |

| NVM | DRAM analogue |
|---|---|
| nvl T* | T* |
| nvl_alloc_nv | malloc |
| *automatic* | free |

# Programming Model: Accessing NVM



```
nvl_heap_t *heap =
  nvl_open("A.nvl");
```

How do we access allocations within an NVM heap?

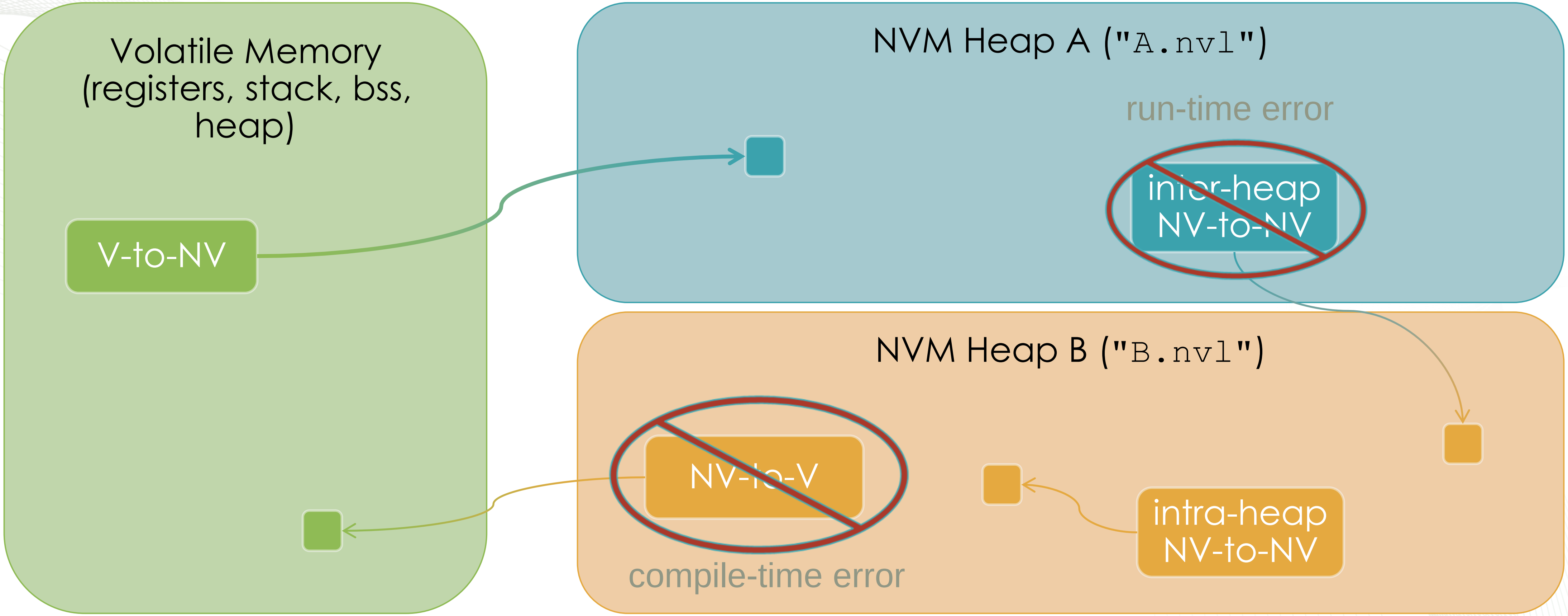
```
nvl T *root =
  nvl_get_root(heap, T);
```

Checksum error if `T` is incorrect type.

Set root with `nvl_set_root`.

Before first `nvl_set_root`, `nvl_get_root` returns null.

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



avoids dangling pointers when memory segments close

Oak Ridge National Laboratory

# Programming Model: Transactions: Purpose

- Ensures data consistency
- Handles unexpected application termination:
  - Hardware failure (e.g., power loss)
  - Application or OS failure (e.g., segmentation fault)
  - NVL-C safety constraint violation (e.g., inter-heap NV-to-NV pointer)
- Does not handle concurrent access to NVM:
  - Future work
  - Concurrency is still possible
  - Programmer must safeguard NVM data from concurrent access

# 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

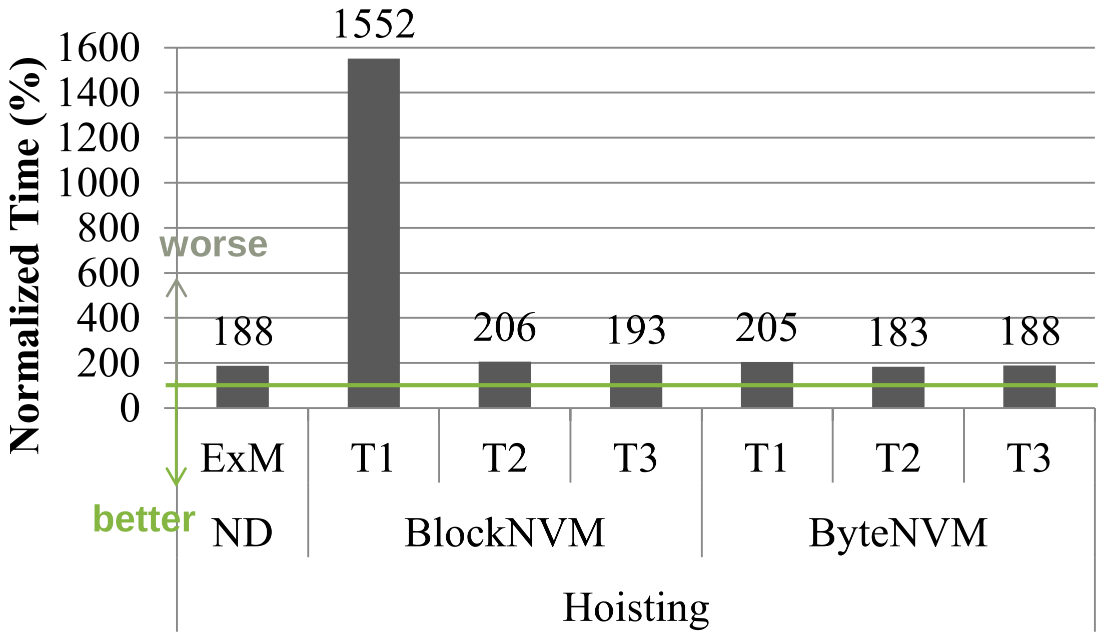
```
#include <nvl.h>
void matmul(nvl float a[I][J],
            nvl float b[I][K],
            nvl float c[K][J],
            nvl int *i)
{
  while (*i<I) {
    #pragma nvl atomic heap(heap)
    {
      for (int j=0; j<J; ++j) {
        float sum = 0.0;
        for (int k=0; k<K; ++k)
         sum +=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

- Atomicity:
  - Incomplete transaction rolled back next time NVM heap is accessed
- Consistency:
  - Transactions begin and end with NVM data is in a consistent state
  - Implicit transactions: specify NVL-C internal data consistency
  - Explicit transactions: specify application data consistency
- Isolation (handles concurrent access):
  - Not guaranteed yet
- Durability:
  - All NVM writes are durable when transaction commits

# Evaluation: MATMUL



- ExM = use SSD as extended DRAM
- T1 = BSR + transactions
- T2 = T1 + `backup` clauses
- T3 = T1 + `clobber` clauses
- BlockNVM = `msync` included
- ByteNVM = `msync` suppressed

- Log aggregation (`backup`) is important for performance
- `msync` is the culprit
- Skipping undo logs (`clobber`) has little to improve upon
- NVL-C has minimal overhead

# 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

# Acknowledgements

- Contributors and Sponsors
  - Future Technologies Group: http://ft.ornl.gov
  - US Department of Energy Office of Science
    - Exascale Computing Project
    - DOE Vancouver Project: https://ft.ornl.gov/trac/vancouver
    - DOE Blackcomb Project: https://ft.ornl.gov/trac/blackcomb
    - SciDAC RAPIDS Project
  - US DARPA

Oak Ridge National Laboratory

# Bonus Material

OAK RIDGE
National Laboratory