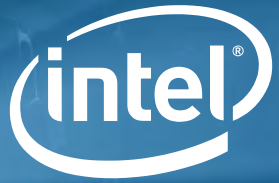
THE INTERTWINED FUTURES OF HIGH PERFORMANCE COMPUTING AND ARTIFICAL INTELLIGENCE



Al Gara
Intel Fellow,
Data Center Group at Intel Corporation

IXPUG
April 23rd 2018

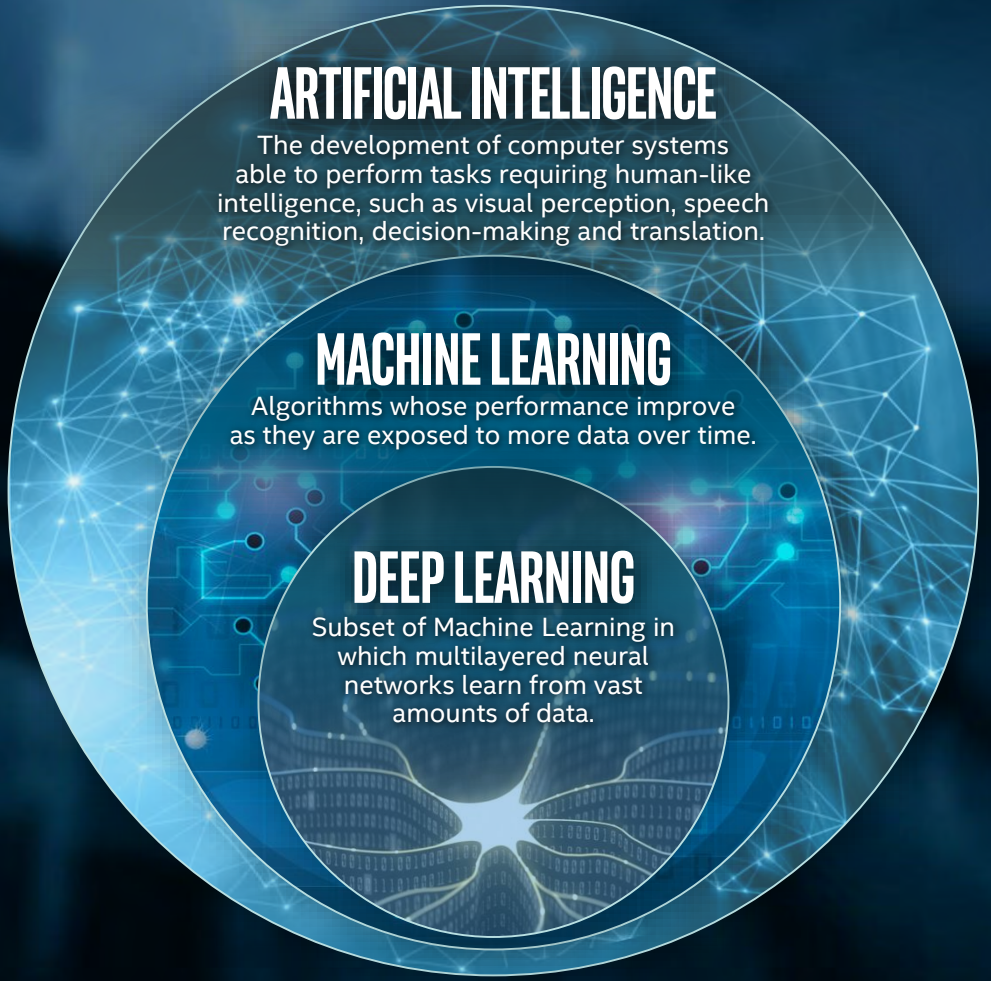
Today:

THE NATURAL CONVERGENCE OF AI AND HPC

TECHNOLOGIES THAT WILL GET US TO EXASCALE CAPABILITIES

10 TO 15 YEAR CRYSTAL BALL LOOK AT SYSTEMS AND TECHNOLOGIES

WHAT IS AI?


ARTIFICIAL INTELLIGENCE
The development of computer systems able to perform tasks requiring human-like intelligence, such as visual perception, speech recognition, decision-making and translation.
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time.
DEEP LEARNING
Subset of Machine Learning in which multilayered neural networks learn from vast amounts of data.

AI ON HPC – UP TO SCALE

U.S. DEPARTMENT OF ENERGY | Office of Science

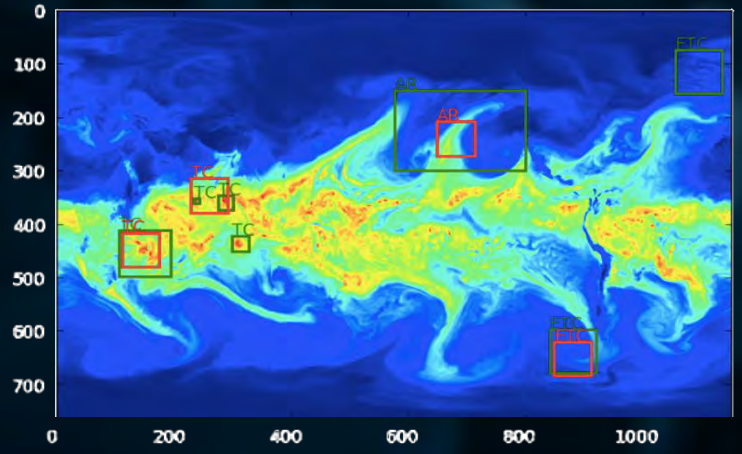
BERKELEY LAB

SCALE MATTERS!

Top Performance for larger datasets and batch sizes

Aggregated processing with Ks nodes

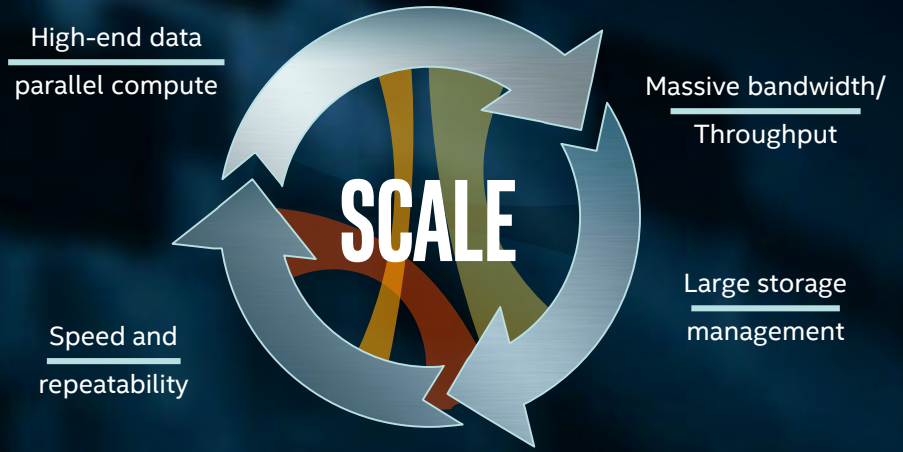
E.g., Climate Pattern Discovery - 15 PF¥



TIMELY ITERATIONS ENLIGHTEN

Time-to-Train requires quick cycles

Experimentation may take multiple tries

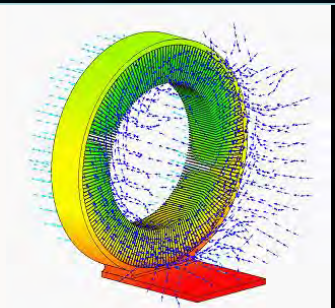


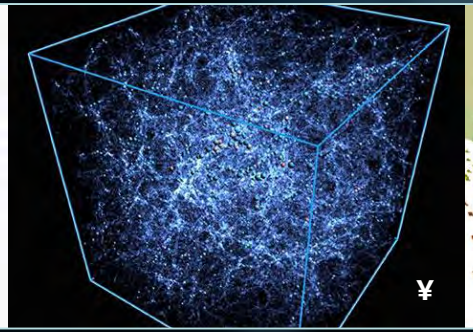
¥ US Department of Energy Office of Science and Berkeley Lab

HPC ON AI – HARNESSING DL

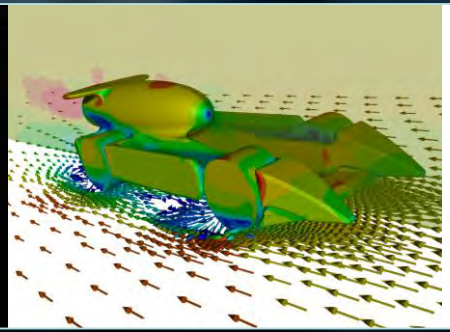
U.S. DEPARTMENT OF ENERGY | Office of Science

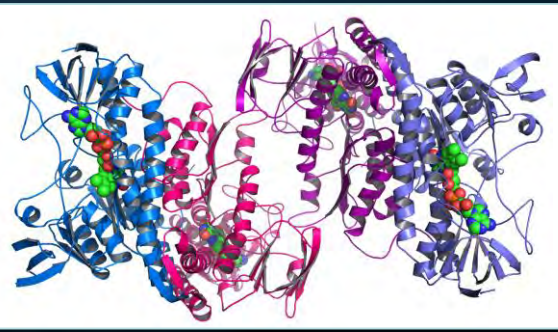
BERKELEY LAB
Bringing science solutions to the world





¥





POWERFUL CAPABILITY FOR HIGH PERF DATA ANALYTICS

COMPLEMENTARY TO MODELING & SIMULATION

No need for complete / complex models

Supervised, Semi-, and Un-supervised

DL METHODOLOGIES & CAPABILITIES : A GREAT MATCH

Pattern Classification, Clustering, Feature Learning, Anomaly Detection

BREAKTHROUGH OPPORTUNITIES

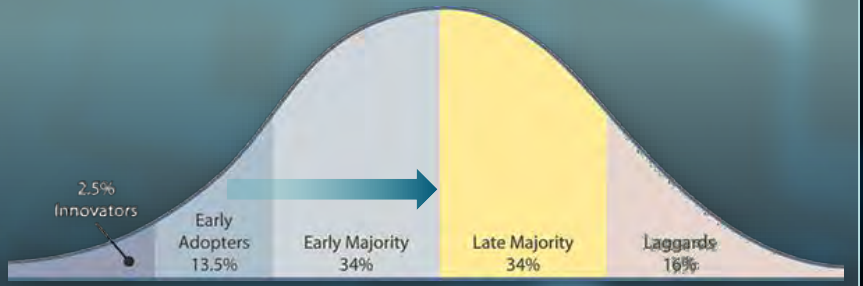
Precision Medicine

'Faint Signal' Fraud Detection

¥ US Department of Energy Office of Science and Berkeley Lab

KEY TRANSITIONS (3-5 YEARS)

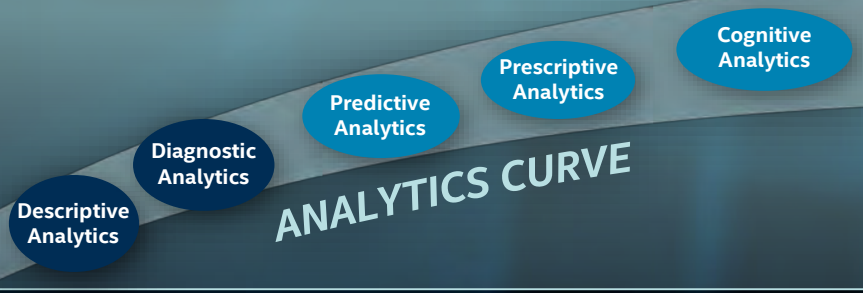
EARLY AI ADOPTERS ➡ LATE MAJORITY



RECOGNITION ➡ IN-CONTEXT 'UNDERSTANDING'

"For sale: baby shoes, never worn."

DIAGNOSTICS ➡ PREDICTIVE ➡ PRESCRIPTIVE ➡ GENERATIVE



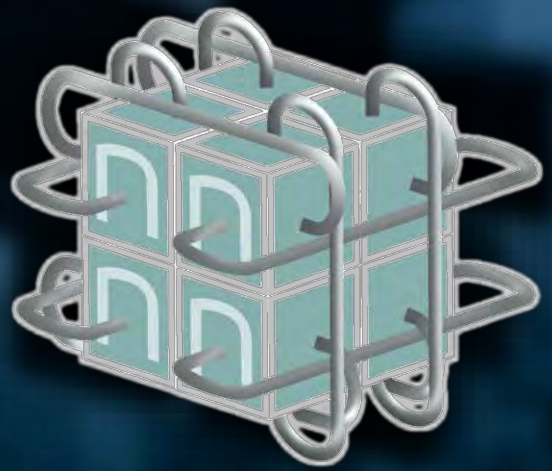
TOPS ➡ PETAOPS ➡ EXASCALE

Extended Performance and Power Efficiency Requirements

DEEP LEARNING – ARCHITECTURAL PERSPECTIVE

Rethinking of compute/memory/interconnect:

- Tensors as native construct
- Spatial processing, highly data-parallel, targeted operations
- Specialized memory hierarchy; high local capacity



- Ultra fast connectivity intra- and inter-die
- Tight SW control over computation, memory and data flow
- Optimal numerics

DEDICATED DL ACCELERATION (DLA): PERFORMANCE & POWER-EFFICIENCY BENEFITS

For heterogeneous end-to-end usages, tight interaction of powerful CPU and DLA will perform best

AI IN THE DATACENTER

Intel® Stratix® 10 FPGA

FLEXIBLE ACCELERATION

Add to accelerate the widest range of AI and other workloads

+

Intel® Xeon® Processor Scalable Family

FOUNDATION OF AI

Begin your journey with the AI you need on the platform you know

+

Intel® Nervana™ Neural Network Processor*

DEEP LEARNING BY DESIGN

Add to accelerate your *intensive* deep learning deployment

†Codename for product that is coming soon
1Formerly codenamed as the Crest Family
All performance positioning claims are relative to other processor technologies in Intel's AI datacenter portfolio
All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

INTERTWINED FUTURES: OPPORTUNITIES AND CHALLENGES

AI DATASETS GROWING TO EXASCALE

- Added Speed and Operability

HIGH-PERFORMANCE DATA ANALYTICS

- Added Flexibility and Efficiency

BREAKTHROUGH COMBINATION

- Unveiling New Technology Opportunities
- Exploring New Technology Challenges

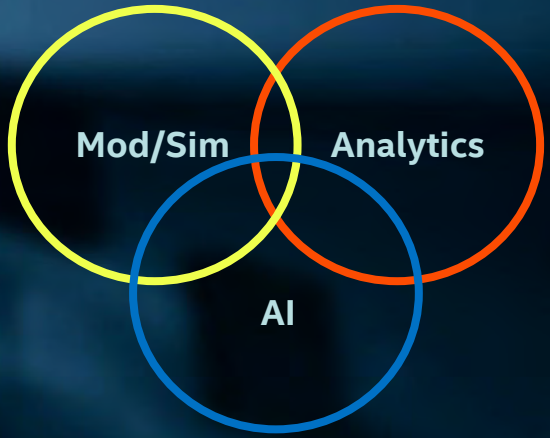
HETERO. WORKLOADS: SIMULATION, DATA ANALYTICS AND AI

Exascale machines must treat all as first class citizens

- This is both a hardware and a software requirement

AI has become a major consumer of computing cycles and it is expected to grow

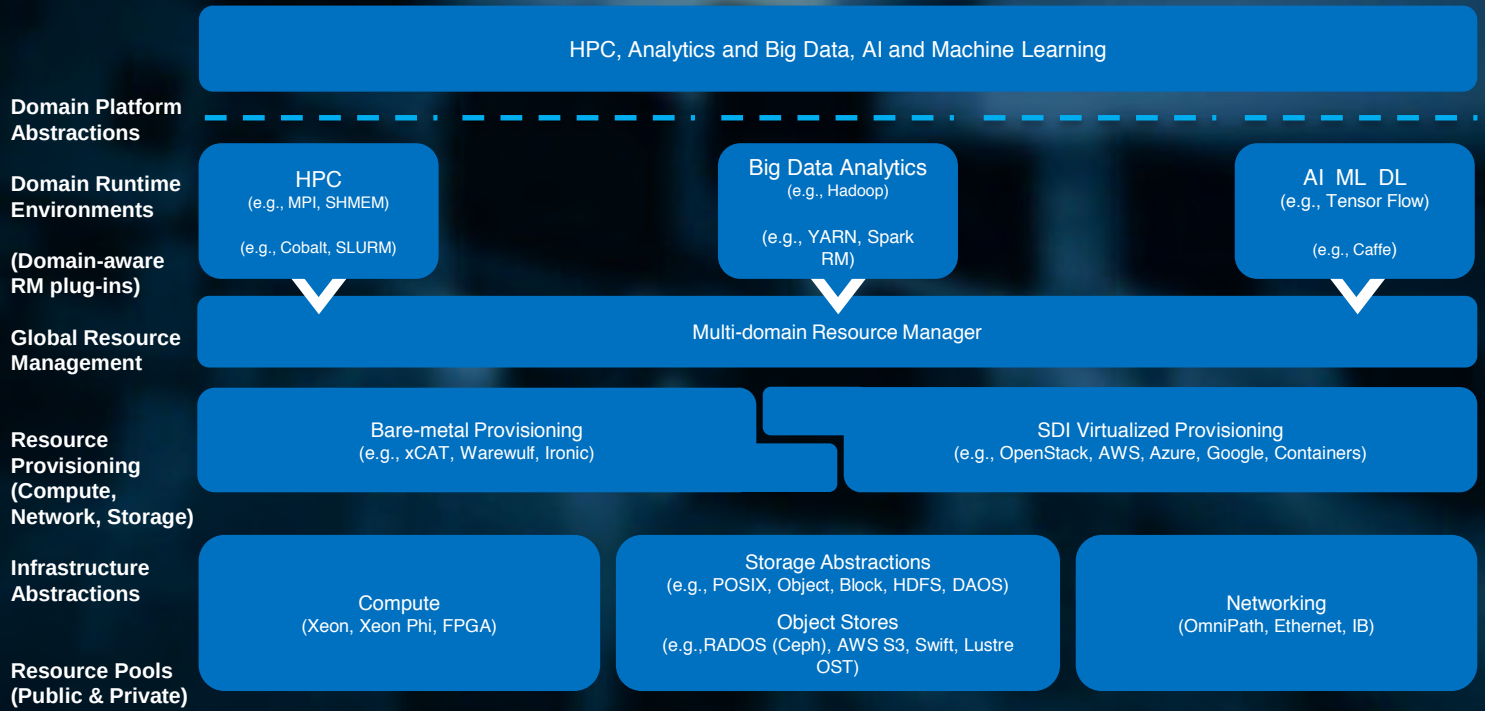
- Compute deployment both at edge and in large cloud
- Will drive economies in fabric, compute with a large focus on power and perf/W



TECHNOLOGIES THAT WILL GET US TO EXASCALE

OPPORTUNITIES AS NEVER BEFORE

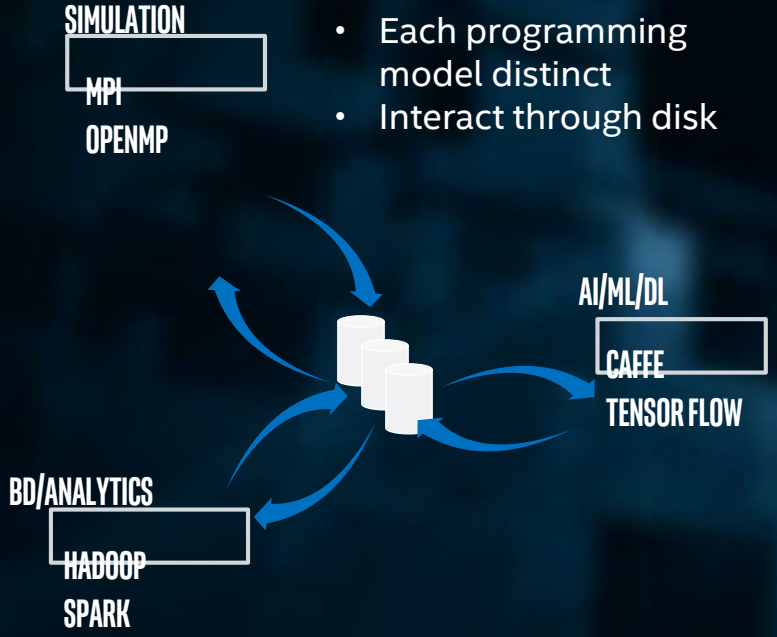
SOFTWARE VISION: UNIFYING THE "3 PILLARS"



VISION: PROGRAMMING THE "3 PILLARS"

TODAY

- Each programming model distinct
- Interact through disk



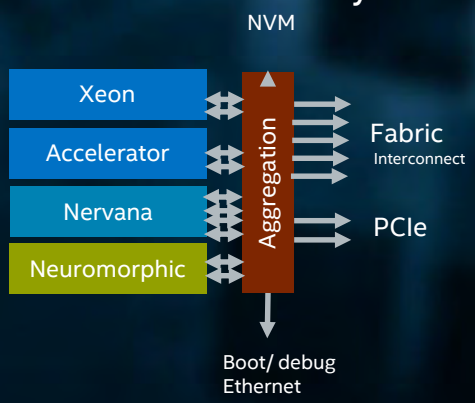
TOMORROW

- Programming models leverage each other
- Interact through memory

FUTURE NODE ARCHITECTURAL DIRECTION

KEY ATTRIBUTE OF SYSTEMS IN THE FUTURE WILL BE HIGHLY CONFIGURABLE AND NEED TO BE ABLE TO SUPPORT UPGRADES TO FUNDAMENTALLY NEW TECHNOLOGIES.

- Many new technologies on the horizon. Want to be able to accommodate when they are available.
- Already see need for larger degree of configurability in systems based on workload focus.
- Same architecture should cover HPC, ML and Data Analytics through configuration.
- Need a consistent software story across these different workloads.



Should allow for tight integration of…

Intel® Xeon Phi™
Intel® Xeon®
Intel® Nervana™
Neuromorphic
FPGA
3DXPT
3DNAND
custom

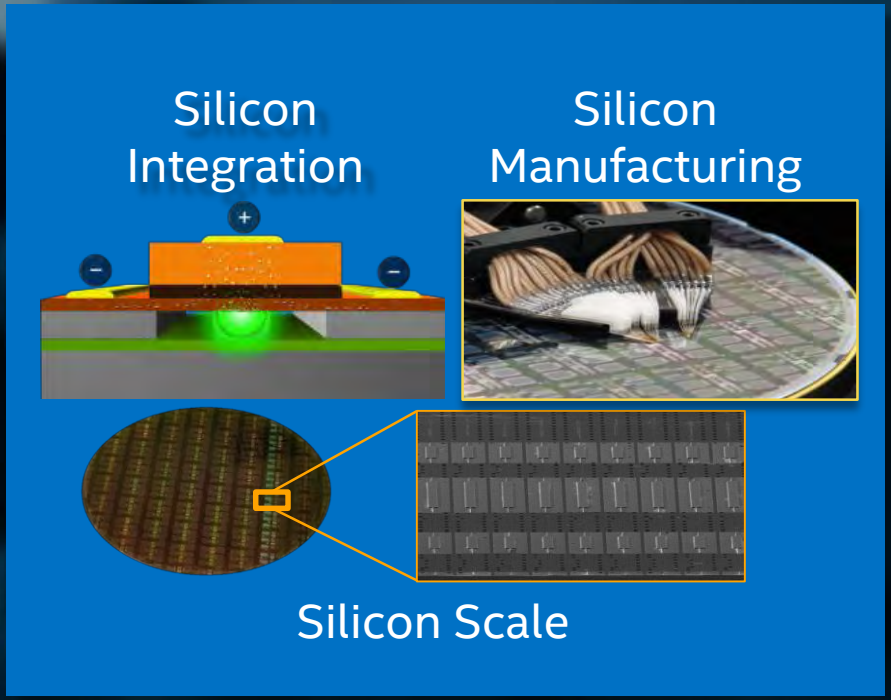
INTEL® SILICON PHOTONICS

OPTICS REPRESENTS ONE OF THE MOST CHALLENGING AND COSTLY ELEMENTS OF A SUPERCOMPUTER.

GAME CHANGING TECHNOLOGIES COMING IN SILICON PHOTONICS.

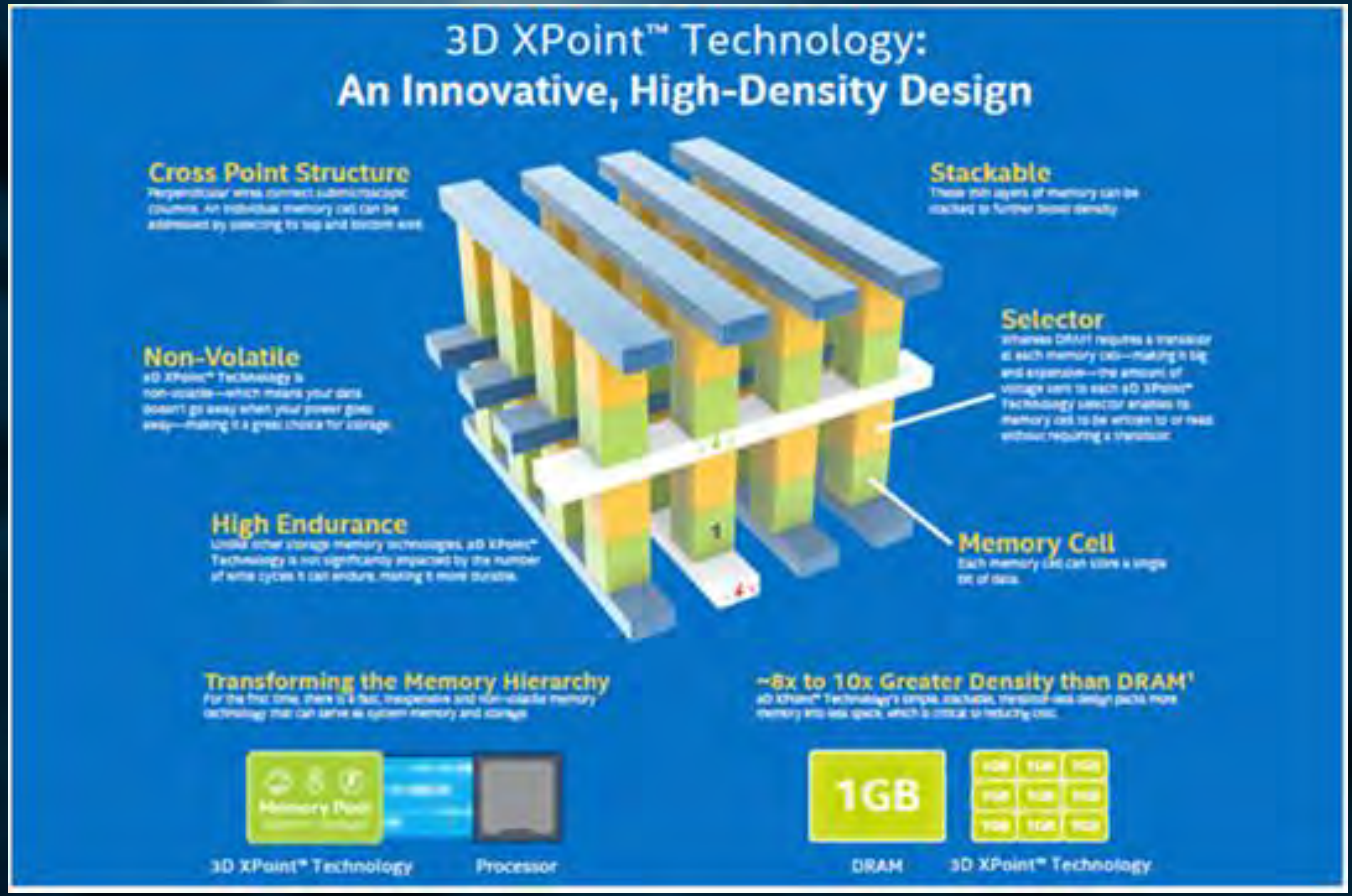
ROOM TO GROW AS WE TRANSITION

- Linear Devices…
- Ring Devices
- Multiple Wavelengths



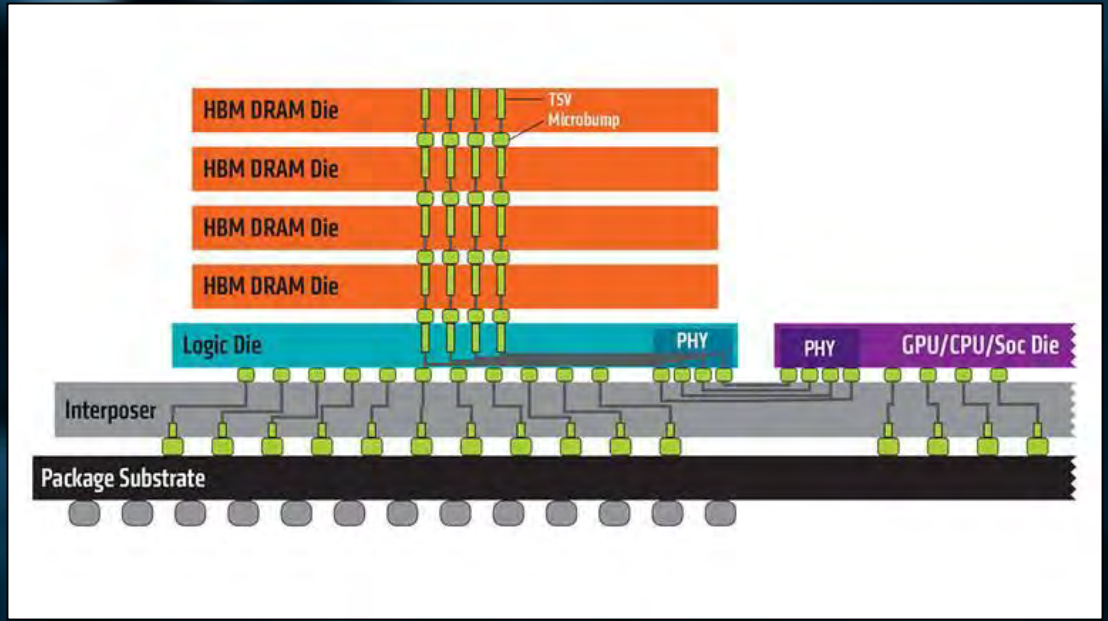
3D XPOINT™

A NEW SUB-CATEGORY OF NON-VOLATILE, FASTER AND DENSER MEMORY/STORAGE TECHNOLOGY

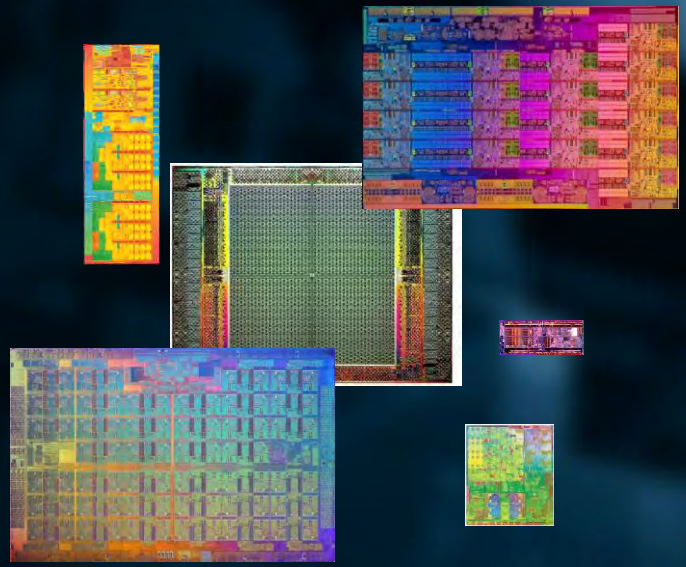


HIGH BANDWIDTH MEMORY (HBM)

- EXTREMELY HIGH MEMORY BANDWIDTH, LOW POWER AND A SMALLER FORM FACTOR
- DRAM IN THE PACKAGE
- CONNECT TO THE SOC BY INTERPOSER
- POSSIBLE USES ARE IN DL, AI, NEURAL NETWORKING, FULLY AUTONOMOUS VEHICLES AND ADVANCED APPLICATIONS THAT REQUIRE SUBSTANTIAL BANDWIDTH AND LOW POWER



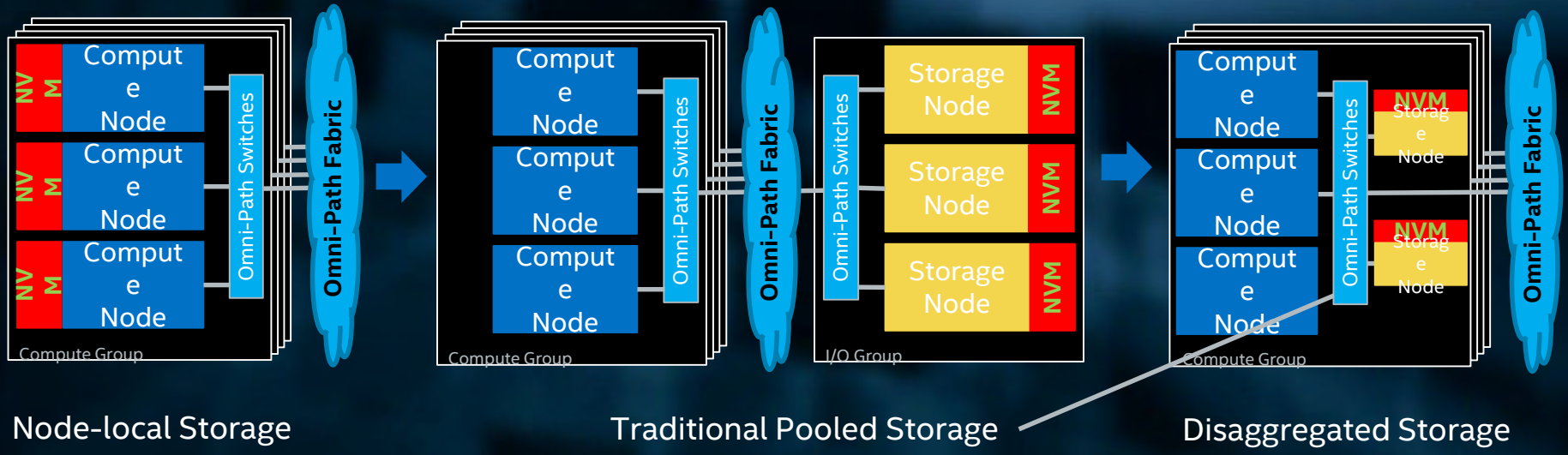
DIVERSITY OF SILICON OFFERINGS



NEW GENERATIONS OF COMPUTE & CAPABILITIES

- The diversity of silicon compute solutions is growing
- Specialized solutions targeted to applications proliferate
- The global demand for processing from the edge to the data center is driving fast growth

STORAGE ARCHITECTURE



Node-local Storage

Traditional Pooled Storage

Disaggregated Storage

NEW STORAGE STACK

EXASCALE STORAGE STACK

- Designed from the ground up for NVM storage
  - Built over **new userspace** NVMe/pmem software stack
  - High **throughput/IOPS** @**arbitrary** alignment/size
  - **Ultra-fine grained** I/O
- Manage **massively** distributed NVM storage
  - **Scalable** communications & I/Os over homogenous, shared-nothing servers
  - **Software**-managed **redundancy**
    - declustered replication & erasure code with self healing

EXASCALE MACHINE: WHAT THIS MEANS FOR USERS

Single unified stack with resource allocation and scheduling across all pillars and ability for frameworks and libraries to seamlessly compose

- Cohesive software stack, Unified control system

Minimize data movement: keep permanent data in the machine via distributed persistent memory and maintain availability requirements

- Enhance with data-aware scheduling

Support standard file-based IO with high performance and provide incremental path to a memory coupling model for highly efficient HPC, BDA, and AI interaction

- DAOS provides high performance file-based and object-based capabilities

Continue driving scalable high performance HPC simulation capability

- Example areas: MPICH, OFI, mOS, optimized libraries

Extreme Performance will be obtainable through high bandwidth memory but capacity will be limited. User focus on enabling more compute per memory capacity will pay dividends

10-15 YEAR CRYSTAL BALL OUTLOOK

Will look at various technologies and the challenges and make an educated guess as to how there will evolve over the next 10-15 years.

- Fabric, Memory, Compute and Power

How much did we know in 2000 as compared to today's reality? (top system grew ~ 50,000x in HPL from then till today)

FABRIC

Foundational to computing at all scale. Currently a cost problem. Power challenges and transients are difficult to handle quickly.

Majority of the cost in the fabric is in the optics for long distance communications.

- Large scale performance for some applications scales with bisection.
- Historically bisection/perf has been falling
- Topology choices will continue to trade off repeatable execution against bisection. Progress has been made in dragonfly-like topologies lately

Silicon photonics will become prime time enabling a steeper drop in $/Gbps

FABRICS: OPTICS REMAINS THE CHALLENGE

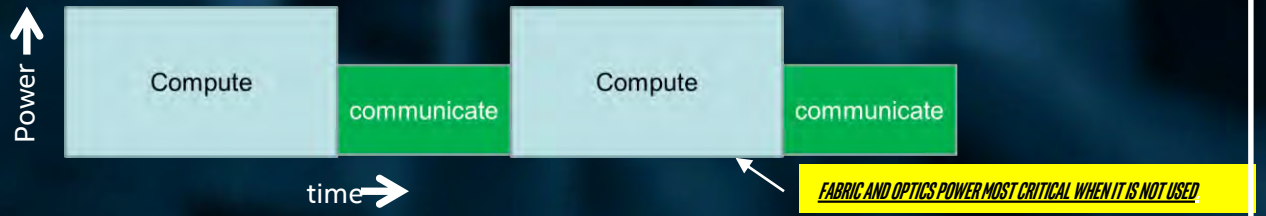
Exascale will not achieve this

IDEAL COST & POWER (REQUIREMENTS @ EXASCALE)

- 3 EXABITS OF TOTAL BW FOR ~3MW AT $30M
- THIS TRANSLATES TO 3M 1TBPS LINKS
  - 1W EACH (1MW/GBPS)
  - $10/LINK OR $0.01/GBPS

20X IMPROVEMENT IN POWER EFFICIENCY
50X IMPROVEMENT IN COST PER BIT

MANY/MOST HIGH END APPLICATIONS ARE TIME STEP BASED



Power and performance trade off. Overlapping will drive up system maximum power or cause performance throttling. Timescale to idle power in fabric will determine whether this is hardware or software.

FABRICS: TOPOLOGY AND OPTICAL BUNDLE SIZE MAY MATTER

Topology	Number of unique Destinations from each rack (N= number of racks)
Dragonfly	~ N^2
Fat Tree	~ N
Torus	4 to 10

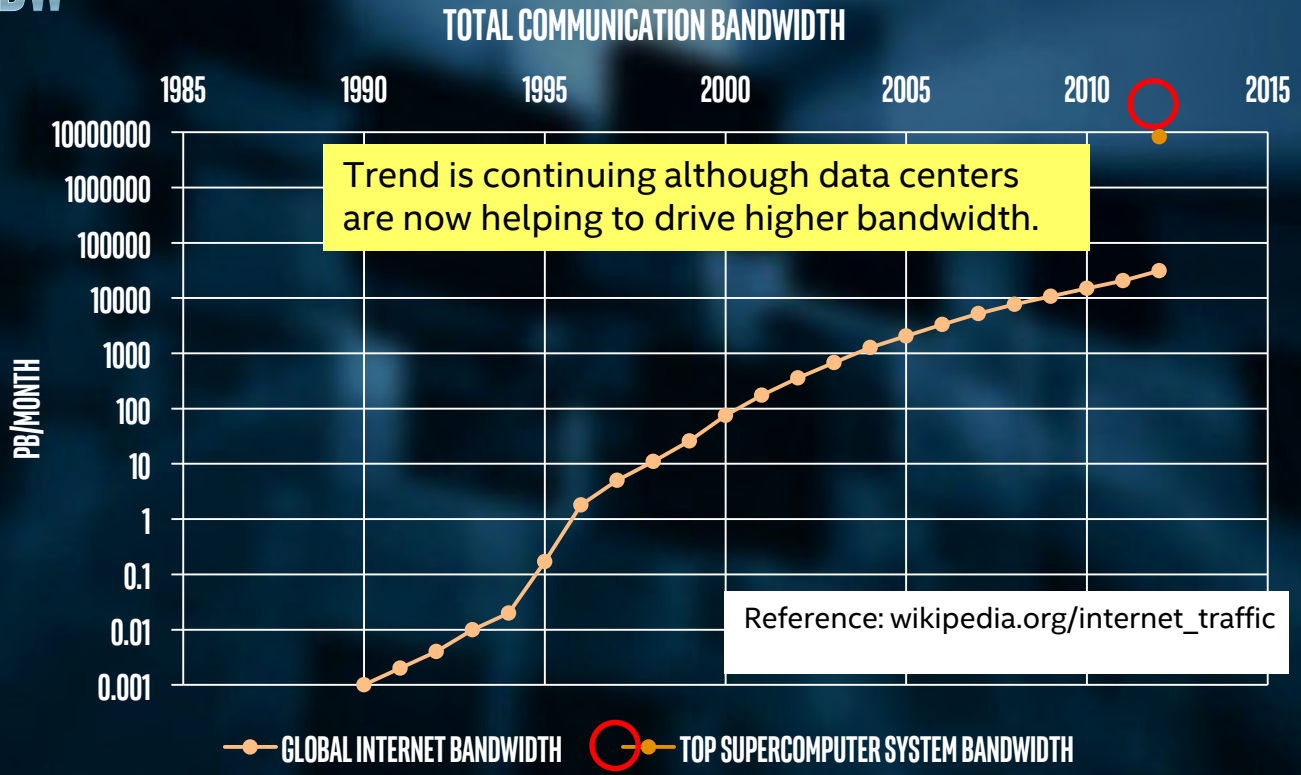
Getting to exceptional cost will require very high BW per fiber to amortize fixed costs connectors, fiber etc.

As BW/$ increases through WDM for example… if $ also raises as BW raises faster this can influence the choice of most cost effective topology.

We are already approaching small numbers of fabric ports in dragonfly topologies for the largest machines.

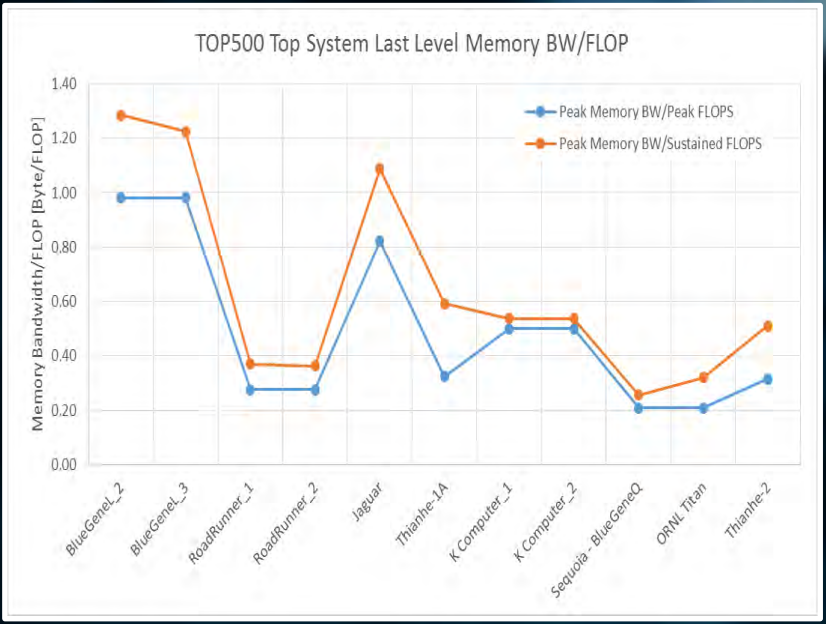
Prognosis: We will take a significant leap forward ~ 2020 (Now looking like 2022) in $/Gbps followed by strong price improvement through 2030. But 1000x would be a big stretch. 20x to 50x more likely.

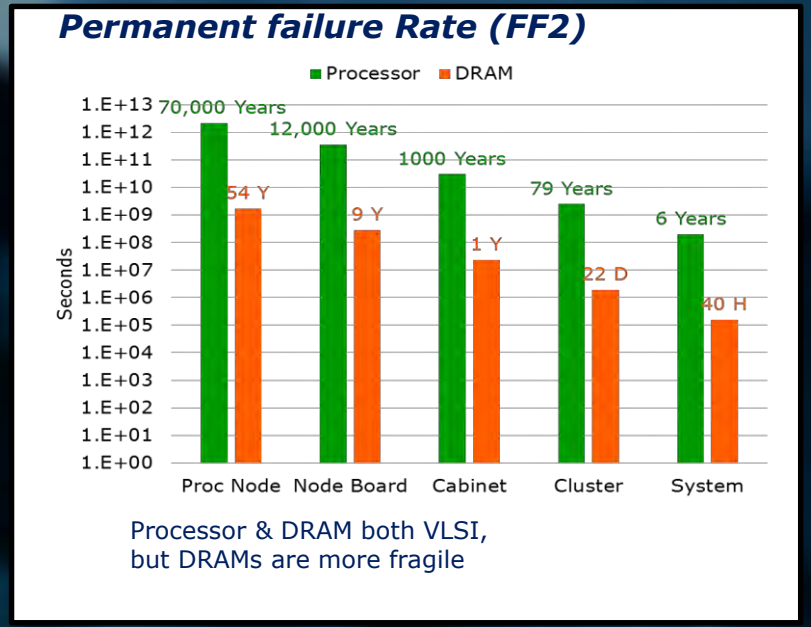
OPTICAL BW OF SINGLE LARGEST SUPERCOMPUTER SWAMPS TOTAL WORLDWIDE INTERNET BW



As data center wakes up to BW needs, we will develop co-travelers in this.

MEMORY: WHERE IS IT GOING…

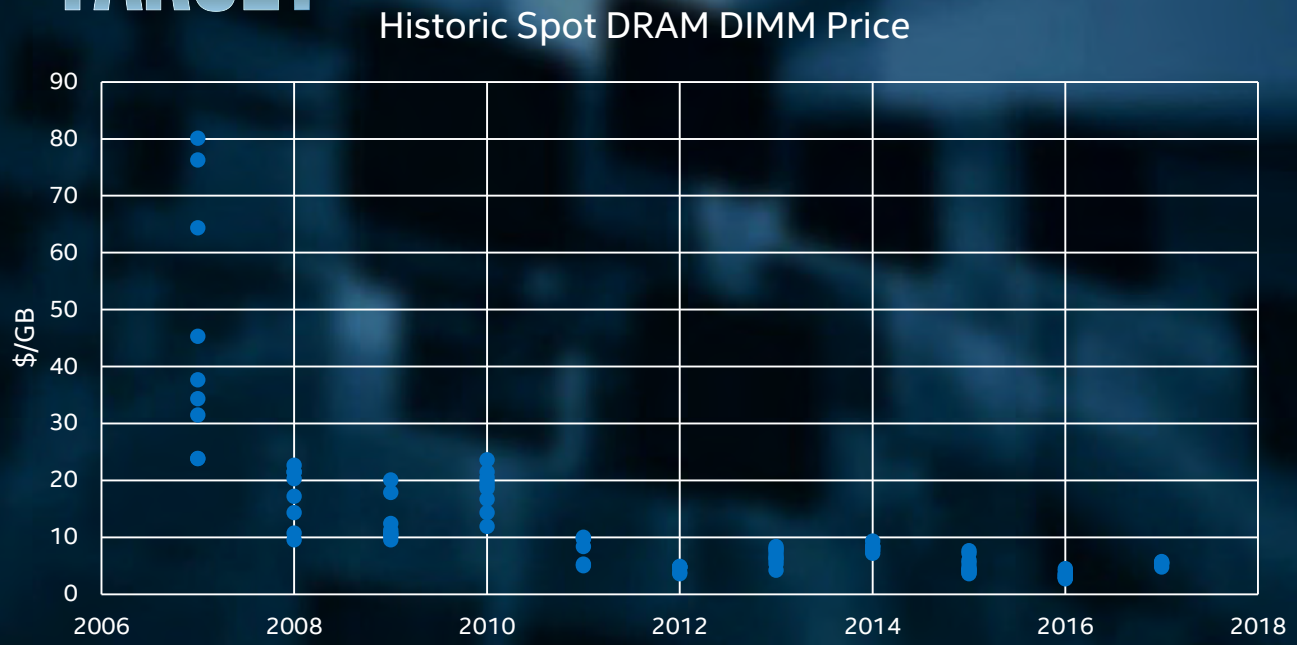




Memory bandwidth has been on a downward trend. Moving to alternatives to DRAM DIMMS is inevitable.

Memory dominates component failure now and into future. Need to carefully consider implications for in-package in this context

MEMORY: ALWAYS A MOVING TARGET



Spot price now is higher than it was in 2012 for commodity DRAM

MEMORY: SOME GOOD, SOME BAD

The Good:

- New packaging and microarchitecture directions have allowed for the future in-package DRAM to keep up with compute
- This looks like a trend we can continue.
- The power improvements are also critical to the future.
- New memory technologies emerging

	DRAM DIMMS	HIGH BW MEMORY
COST/ CAPACITY	1X	1.2X TO 1.5X (SHORT TERM VIEW IS ~ 3X)
COST/BW	1X	1/10X TO 1/40X (NOW 1/5 TO 1/20)
POWER/BIT	1X	1/2 X TO 1/10 X

The Bad:

- The capacity story far less compelling. Slowing improvements in capacity/$
- New emerging technologies are chasing a moving target in terms of BW. Expect a long time before they catch in-package DRAM
- Very good news for I/O in the future with NVMEM running at ~ 1/10 in terms of $/capacity.

The Prognosis :

- Will have a compelling BW story through 2030 but will struggle with slow capacity growth
- Will need to continue to get more performance out of same capacity gen over gen. (Threads etc)
- New memory technology will start further away and move closer but not likely to replace DRAM cells by 2030

PROCESSOR: SOME GOOD SOME BAD, SOME DISRUPTIVE

The Good:

- HPC is a fast growing business strategic to both commerce and governments.
- Investments in HPC including processor development have been strong and appear to be accelerating.
- This will enable innovation at the processor level in ways that we did not have on our radar screen 5 years ago.
- Silicon scaling is still moving albeit at a slowing pace. It always accounted for at most half the performance improvement over time.

The Bad:

- Getting even 1000x HPL by 2030 at same power and cost hard to see path to that with silicon scaling slowing. Things look much better for general purpose perf.
- All paths lead to exploiting concurrency new ways. Frequencies not going up any time soon.

The Disruptive :

- New forms of compute could dramatically change the game for some algorithms.

PROCESSOR: THE DISRUPTIVE POTENTIAL

Recent explosion in exploring new forms of compute:

- Neuromorphic is probably most promising but...
  -- still working to find broad class of algorithms.
- Activities into many other areas.. Quantum, low bit arithmetic, information encoding,
- Two paths for technology...

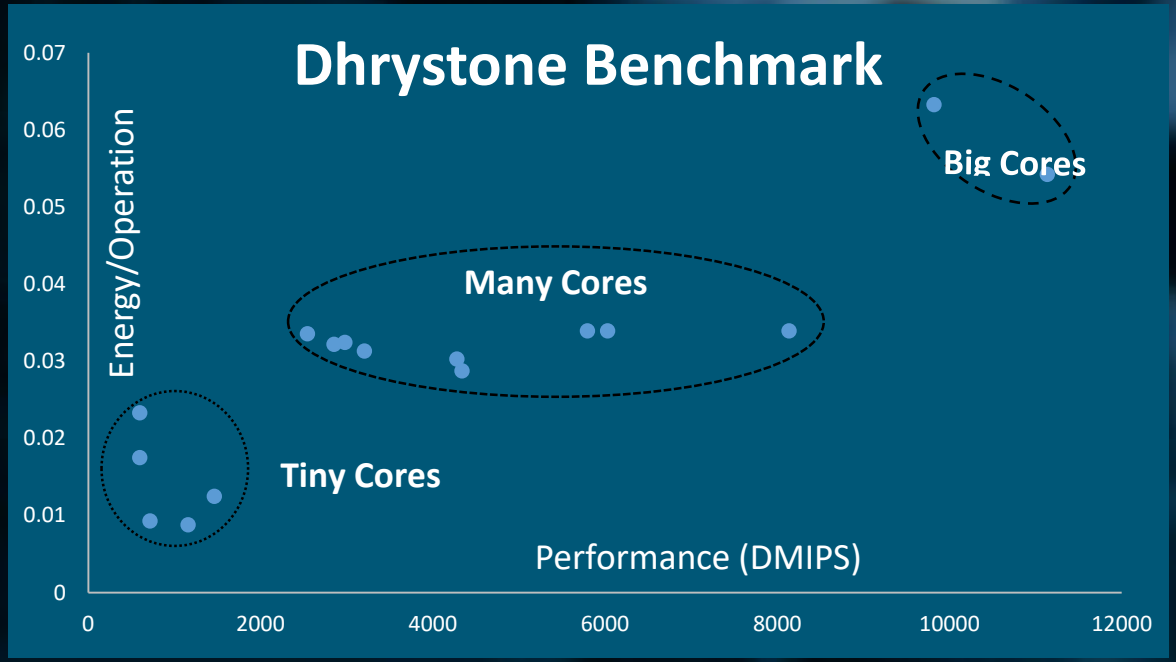
  1) Special purpose device
  2) Augmentation of “traditional compute”
- New forms of compute offer disruptions in both performance and energy efficiency.
- How perf/W tracks will be critical to integration into conventional systems. Can’t have an accelerator in a general purpose machine that drives 10x the power when it is used. That is a special purpose machine.
- Will start to see new technologies and techniques that will be focused on information processing per watt (backup talk)

The Prognosis :

- New forms of compute will dramatically change the game for some algorithms. Expect that integration into traditional systems will be fastest path to growth of new technologies. This will require some compromises and the merging of very different technologies can be difficult.
- Quantum and Neuromorphic will likely be used as special purpose accelerators before 2030.

POWER CHALLENGES

EFFICIENCY OF VARIOUS CORES (HARDWARE)

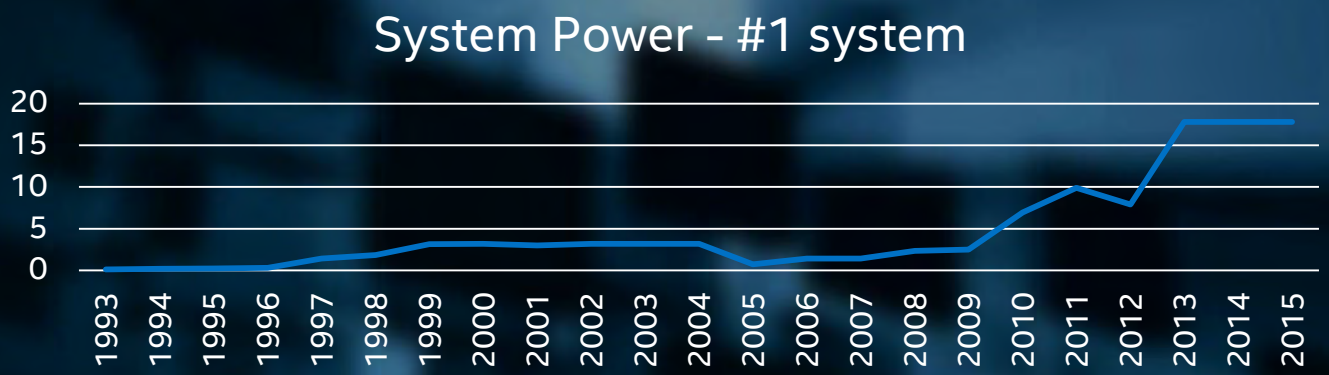


Performance has a clear inverse relationship with energy efficiency at the single core level.

The situation becomes much less clear at the application level.

Source: Intel's internal evaluation for the existing CPUs. The Dhrystone benchmark is publicly available.
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit: http://www.intel.com/performance. Source: Intel measured or estimated as of November 2017

POWER: THE LONG TERM CHALLENGE



We have had a recent growth in power. Where will this really go in 2030?

Looking at this from a simple TCO perspective….

Using $1M/MW-year current rule of thumb.
Assume this drops to $0.5M/MW-year in 2030
If we want to spend ~ 50% of the system total cost on power over 5 years then ~$200M machine cost = 80MW * 5 years * $0.5M/MW-year

Evolution of Systems Over Time
Crystal ball version…

Power will REALLY be the challenge.

	SEQUOIA	2018	2022	2026	2030
PEAK FLOPS	20 PF	~200 PF	2 EF	20 EF	100 EF
PERF (GENERAL)	1X	10X	100X	1000X	5,000X
PERF SPECIALIZED	1X	10X	200X	5000X	50,000X
MEMORY CAP (HIGH PERFORMANCE)	1.6P PB	2.5 PB	8 PB	30 PB	60 PB
MEM BW	2.5 PB/S	30 PB/S	200 PB/S	1500 PB/S	5,000 PB/S
FAB BISECTION	50 TB	1 PB/S	10 PB/S	50 PB/S	100 PB/S
FAB INJECTION	2 PB/S	3 PB/S	30 PB/S	150 PB/S	300 PB/S
POWER	8 MW	18MW	25MW (PROBABLY 40MW NOW)	50MW (65 MW?)	80MW (100MW?)

Requires system arch disaggregation

With growth in power we will see continued growth in systems for some time.
Expect a financial power cap to hit before 2030. Considerable slowing after that time.
500x to 1000x peak through 2030. (as compared to our previous 50,000x)
General purpose will be closer to 5000x

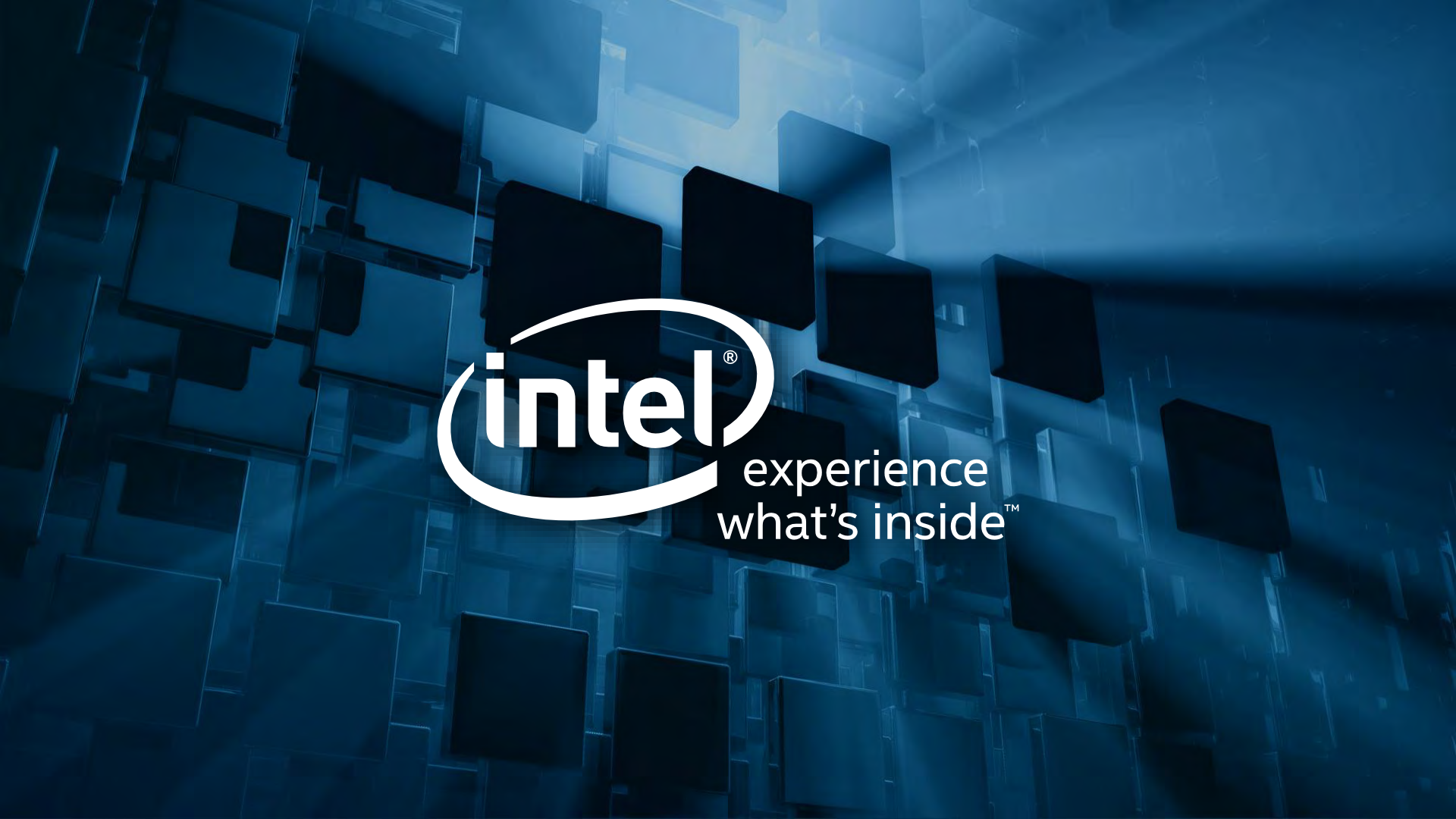
CONCLUSION

- New and emerging technologies are enabling new building blocks from which we will architect future systems
- Future systems will enable both excellent legacy performance and a transitional path to exascale performance.
- User focus on computation per memory will pay big dividends across architectures.
- We have an exciting path to exascale which will mean multi-PF racks for broad HPC usage
- System Software and User applications will continue to have opportunity to stay on the exponential performance growth curve through exascale and beyond.
- Power of systems is an increasing challenge. More focus on TCO optimization.
- The convergence of AI, data analytics and traditional simulation will result in systems with broader capabilities and configurability as well as cross pollination.

LEGAL INFORMATION

This presentation contains the general insights and opinions of Intel Corporation (“Intel”). The information in this presentation is provided for information only and is not to be relied upon for any other purpose than educational. Statements in this document that refer to Intel’s plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel’s results and plans is included in Intel’s SEC filings, including the annual report on Form 10-K.

Any forecasts of goods and services needed for Intel’s operations are provided for discussion purposes only. Intel will have no liability to make any purchase in connection with forecasts published in this document. Intel accepts no duty to update this presentation based on more current information. Intel is not liable for any damages, direct or indirect, consequential or otherwise, that may arise, directly or indirectly, from the use or misuse of the information in this presentation. Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.


intel®
experience
what's inside™