Criteria for a Scalable Architecture

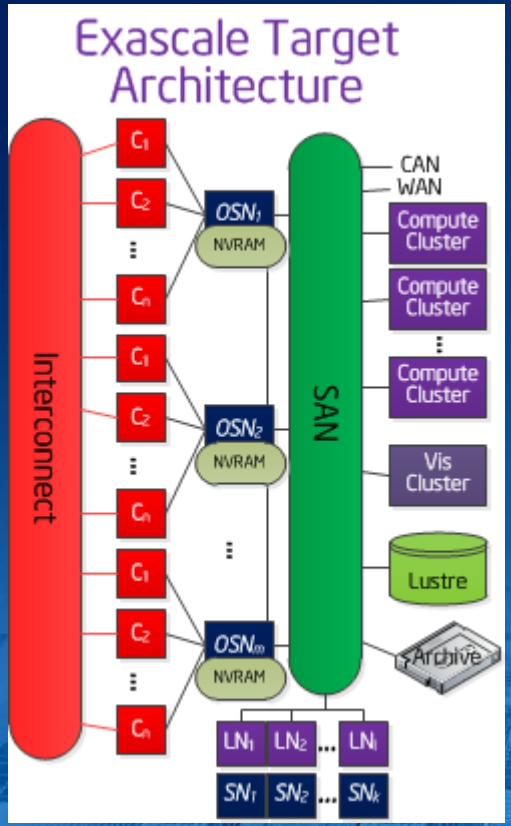
2013 OFA Developer Workshop, Monterey, CA

Mark Seager

CTO for the HPC Ecosystem
Intel Technical Computing Group

OPENFABRICS ALLIANCE

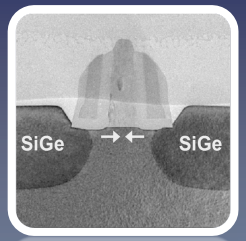
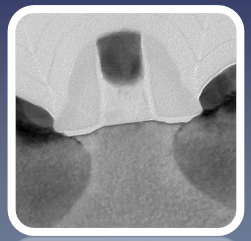
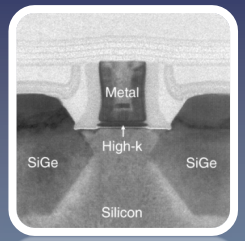
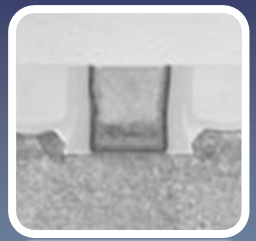
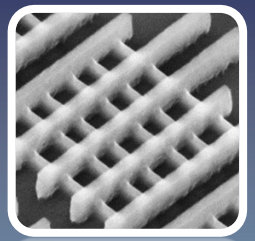
Exascale Systems Challenges are both Interconnect and SAN



- Design with system focus that enables end-user applications
- Scalable hardware
  - Simple, Hierarchal
  - New storage hierarchy with NVRAM
- Scalable Software
  - Factor and solve
  - Hierarchal with function shipping
- Scalable Apps
  - Asynchronous coms and IO
  - In-situ, in-transit and post processing/visualization

intel

Myth: Moore's Law is dead!
Reality - Moore's Law is Alive and Well

2003	2005	2007	2009	2011
90 nm	65 nm	45 nm	32 nm	22 nm
				
Invented SiGe Strained Silicon	2nd Gen. SiGe Strained Silicon	Invented Gate-Last High-k Metal Gate	2nd Gen. Gate-Last High-k Metal Gate	First to Implement Tri-Gate

STRAINED SILICON

HIGH-k METAL GATE

TRI-GATE

The foundation for all computing

22nm
A Revolutionary Leap in Process Technology

37%
Performance Gain at Low Voltage*

>50%
Active Power Reduction at Constant Performance*

intel

LINPAC Performance 2x/1yr CAGR

Performance CAGR driven in part by geometric increase in node counts.
Technology in TOP10 impacts *entire* TOP500 in 6-8 years

123 PFlop/s
17.6 PFlop/s
SUM
N=1
60.8 TFlop/s
6-8 years
1.17 TFlop/s
N=500
59.7 GFlop/s
Jack's Laptop (12 Gflop/s)
iPad2 & iPhone 4s (1.02 Gflop/s)
400 MFlop/s

10 Pflop/s
1 Pflop/s
100 Tflop/s
10 Tflop/s
1 Tflop/s
100 Gflop/s
10 Gflop/s
1 Gflop/s
100 Mflop/s

1993 1995 1997 1999 2001 2003 2005 2007 2009

intel

Slide courtesy Jack Dongarra

Tera→Peta-Scale trends are not sustainable

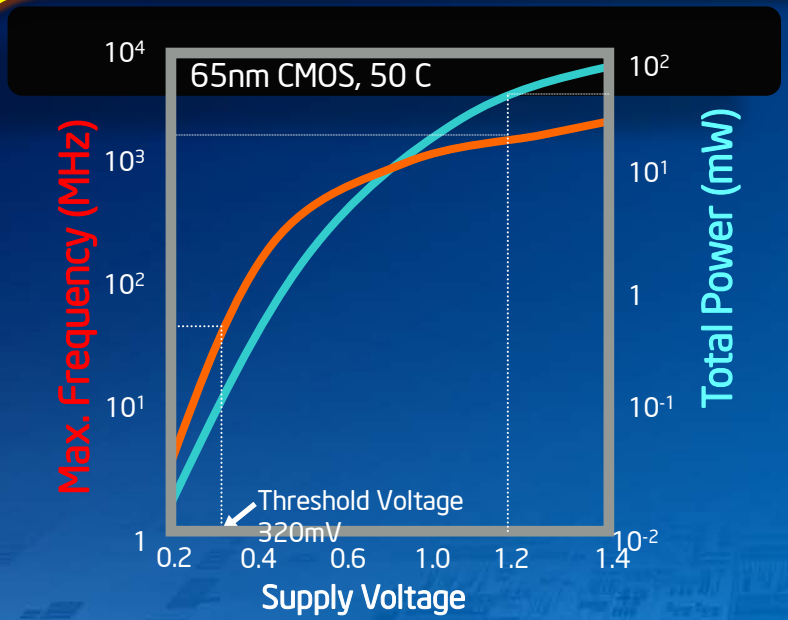
System	date	peak	nodes	cores	power	Facilities Impact
BluePacific ID	1996	0.14	512	512	0.13	B113 Air Handler
BluePacfic TR	1997	1.81	674	2,696	0.25	
BluePacific SST	1998	3.90	1,452	5,808	0.43	B113 2x to 1.8MW
White	2000	12.30	512	8,192	1.00	B451 2x to 3.9MW
BlueGene/L	2004	367.00	65,536	131,072	1.80	
Purple	2005	100.00	1,536	12,288	4.80	New Building 3x
Dawn	2009	501.35	36,864	147,456	1.15	
Sequoia	2011	20,133	98,304	1,572,864	8.00	B453 2x to 30MW

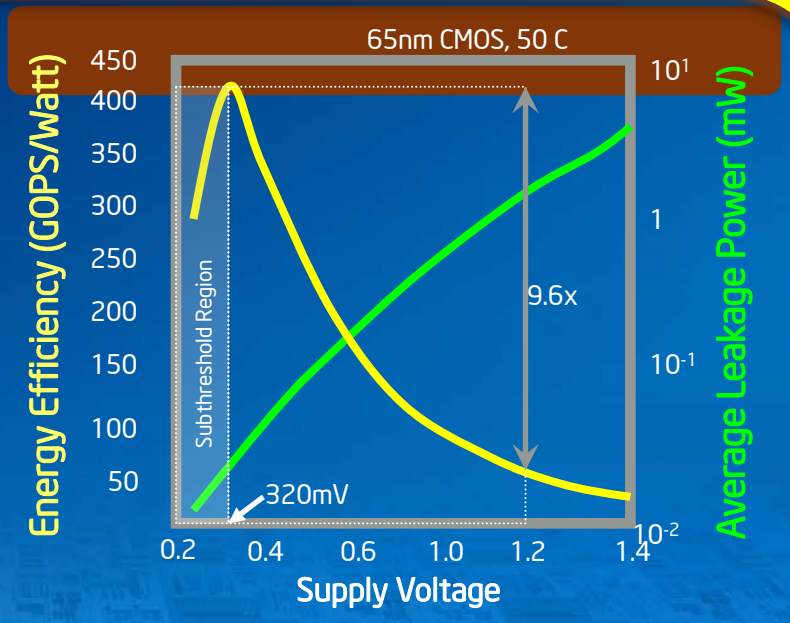
Courtesy LLNL

Where technology will get us…

- First: Continued advances in silicon technology…
  - 2x density every generation. This is half of the story as we have a new silicon generation ~ 2 years.
  - The other factor of 2 every two years has come in through increasing levels of integration and architectural discontinuities.
    - In reality progress is comprised of "evolutionary" curves on top of discontinuous architectural change. Expect another discontinuity around exascale as evolutionary many-core will start to fall off the curve.
- Fundamental changes in memory technology will change the I/O story through integrated NVM

Dealing With Power and Scaling Issues





Low voltage puts pressure on intra-process scaling and cost

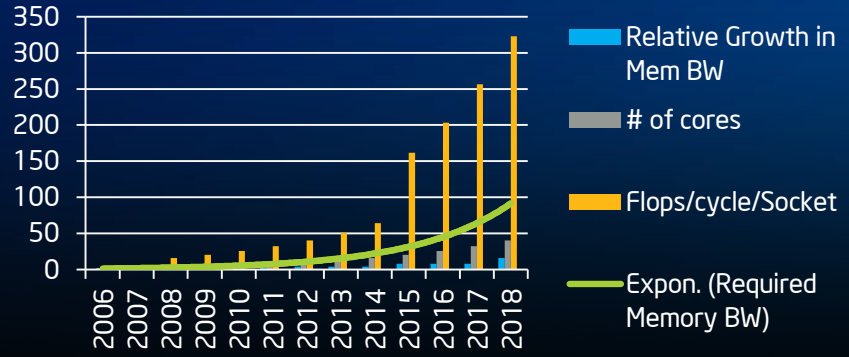
Source: Exascale Computing Study: Technology Challenges in achieving Exascale Systems (2008)

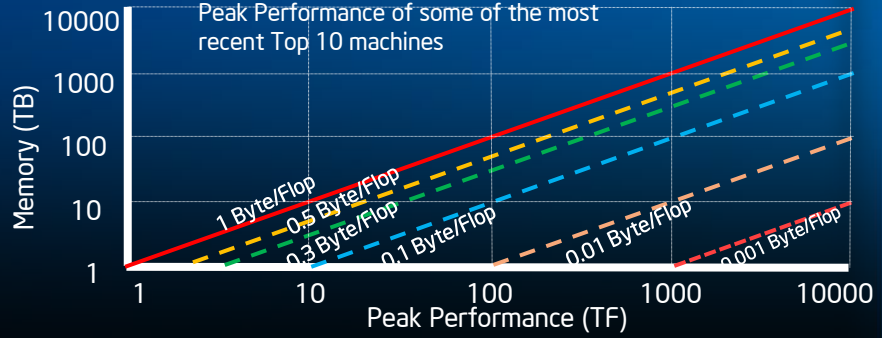
Two key areas of integration are next…

- Memory:
  - New bandwidth/density ratio. Needed to compensate for compute increasing faster than DRAM density. Does fundamentally change the game.
    - Bandwidth is more fundamental than density.
  - New technologies are nearing prime time. Could fundamentally change the game.
- Network:
  - Time to integrate.
    - Performance, cost and power all contribute.

The Memory Challenge to Exascale

Memory Bandwidth and Capacity





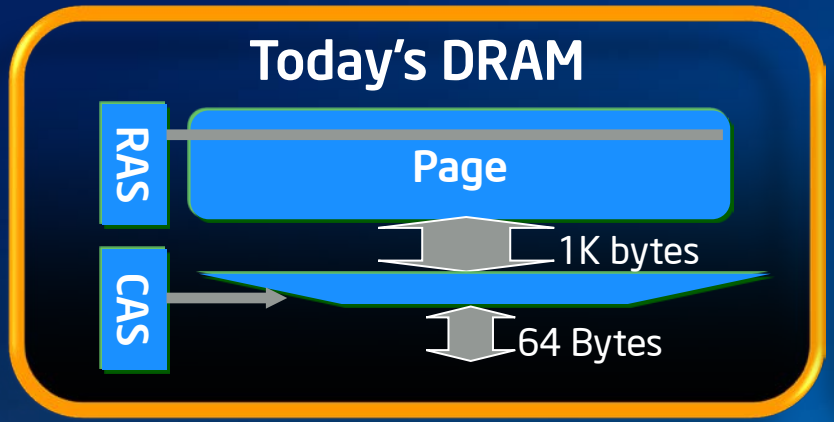
B:F at main memory

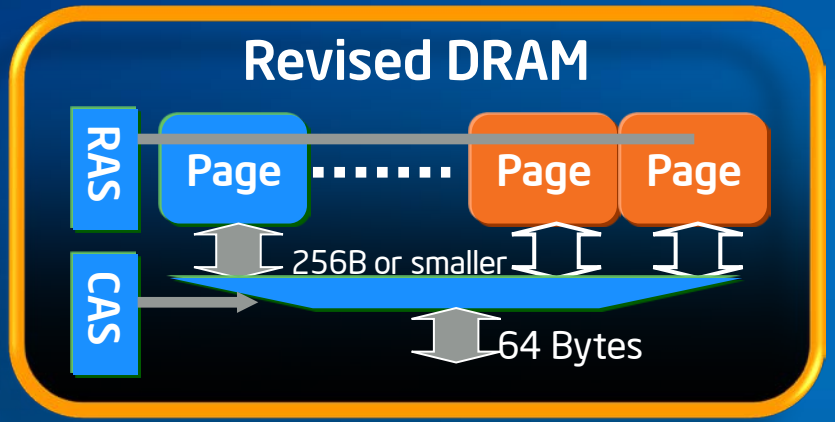
- This was a useful metric before compute got cheap and vector units proliferated.
  - Fundamentally this is the ratio of a very expensive resource divided by a very inexpensive resource.
  - 1.0 B:F for scalar floating (FMA included) still looks right.
  - This metric gets confused on wide vectors (need full scatter gather to get broad multiplicative performance impact).
  - Cache blocking for codes that get high fraction of peak reduces the main memory bandwidth requirements.
  - Yes more is better.

Barriers to innovations sometimes come from the past..

- There are a number of metrics which are commonly used but often not meaningful without context. Need to remove barriers to innovation driven by adherence to legacy metrics.
  - B:F at main store ? (which flops?)
  - B:F at network ?
  - Memory/core ? (threading makes this very weak)
  - Perf/Socket ?
  - What is our measure of the goodness of a machine?

Re-think DRAM Architectures

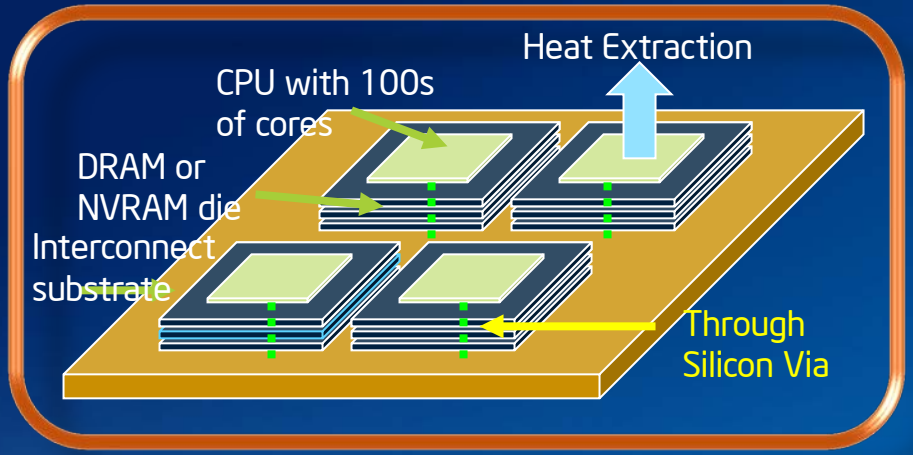


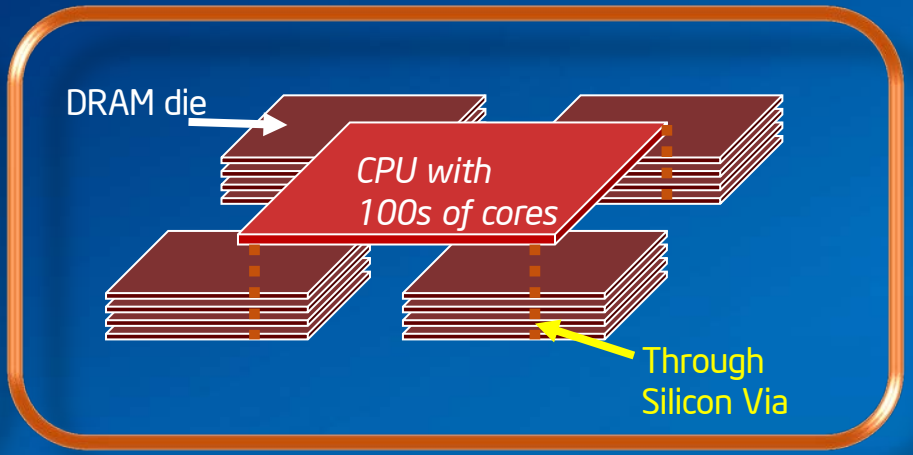


- Activates one large page
- Lots of reads and writes (refresh)
- Small amount of read data is used
- Power wasted in maintaining/ accessing the array

- Activates one smaller page
- Fewer Read and write (refresh)
- Most of the read data is used
- IO can be widened to increase BW

Innovative Packaging & IO Solutions





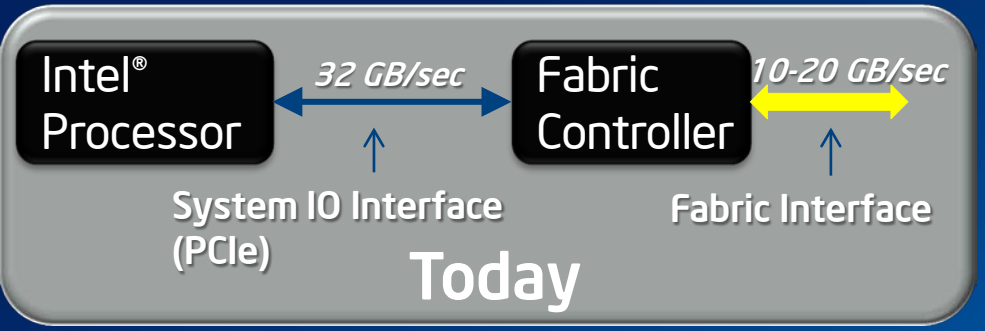
Pins required + IO Power limits the use of traditional packaging

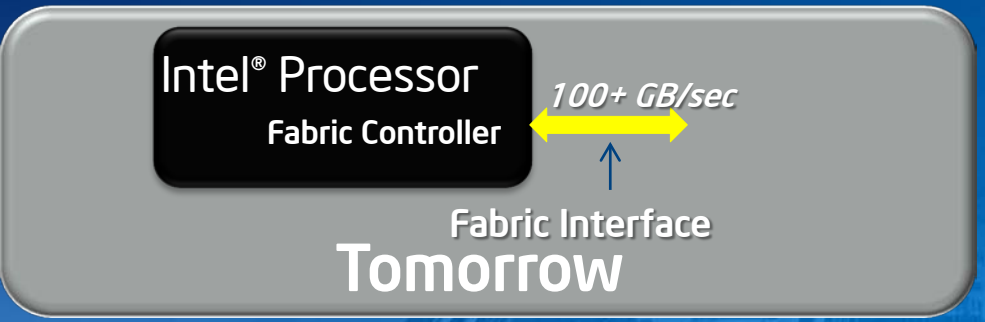
Tighter integration between memory and CPU

High BW and low latency using memory locality

Source: Exascale Computing Study: Technology Challenges in achieving Exascale Systems (2008)

The Advantages of Fabrics Integration

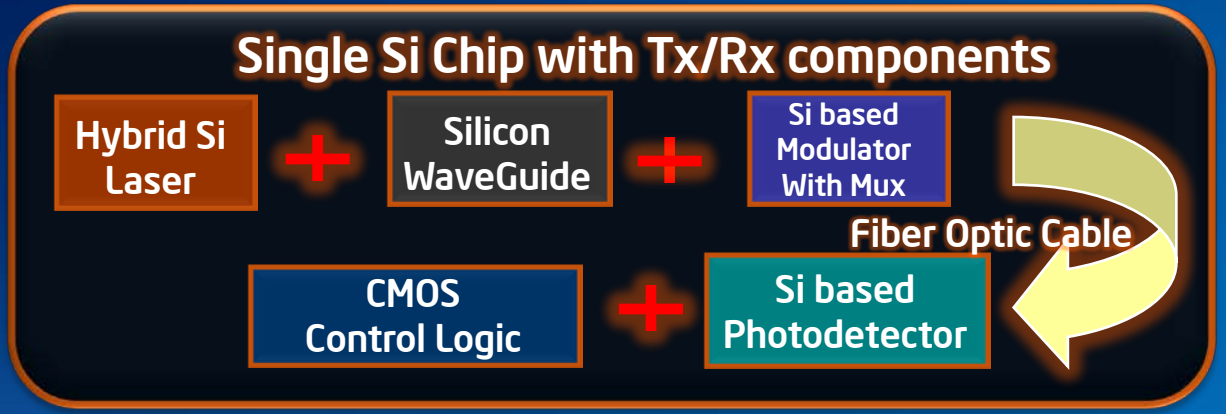




- Problem:
  - Power – System IO Interface Adds "10s Of Watts" Incremental Power
  - Cost & Density – More Components On A Server Node
  - Scalability – Processor Capacity & Fabric Bandwidth Scaling Faster Than System IO Bandwidth
- Solution:
  - Removing The System IO Interface From The Fabrics Solution *Reducing Power*
  - An Integrated Fabrics Results In *Fewer Components On The Server Node*
  - An Integrated Fabric *Balances Fabric and Compute, Scaling Application Performance & Efficiency*

Fabrics Integration Required to Scale Performance & Power

Silicon Photonics

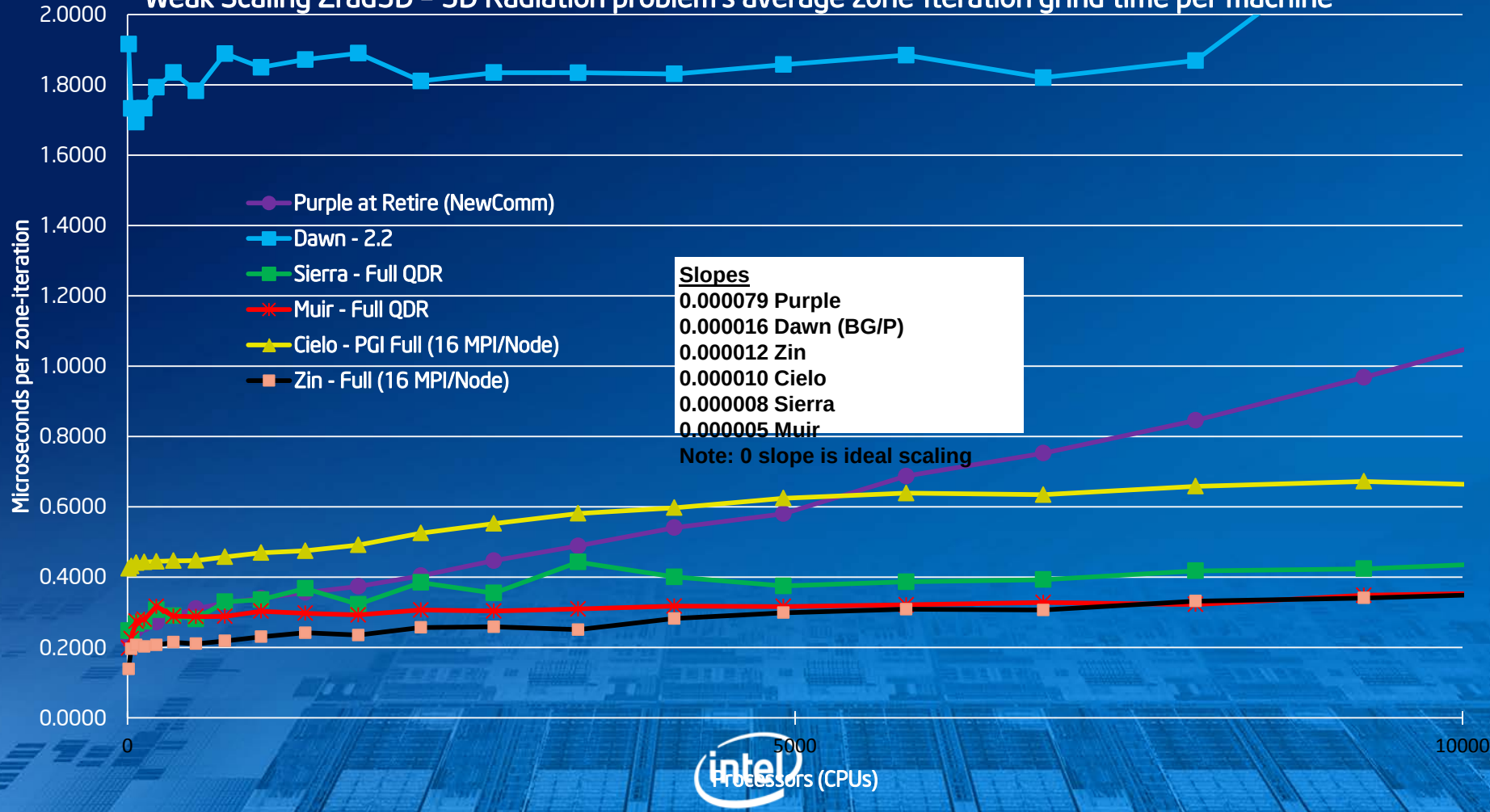


A game changer for long links => Reduces power, latency, and cost
Data delivery over large distances with no EMI effects and high wiring density
Current research shows data transmission rates of >10Gb/s

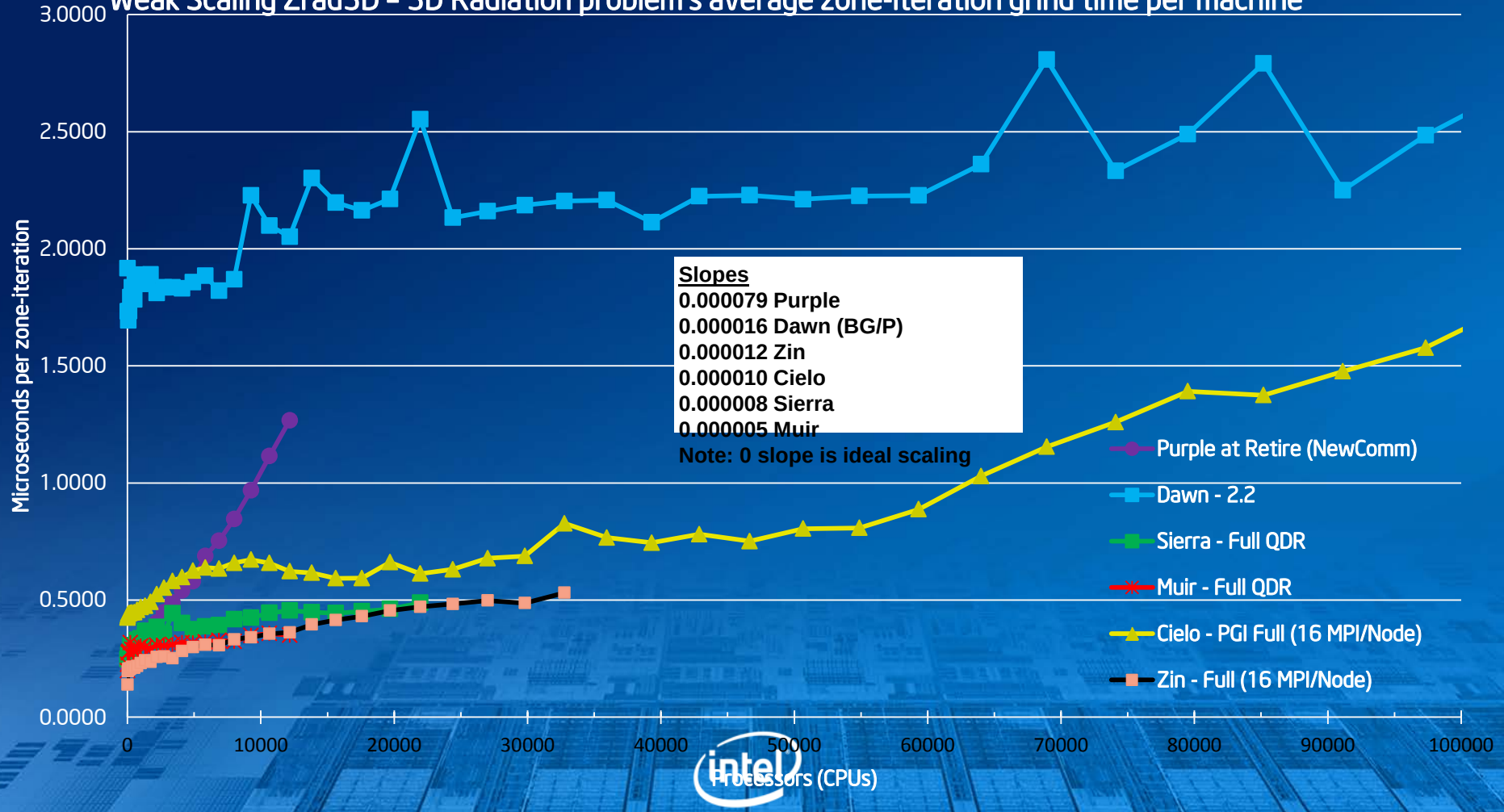
Si Photonics is the Only Solution for Long Links

Source: Intel

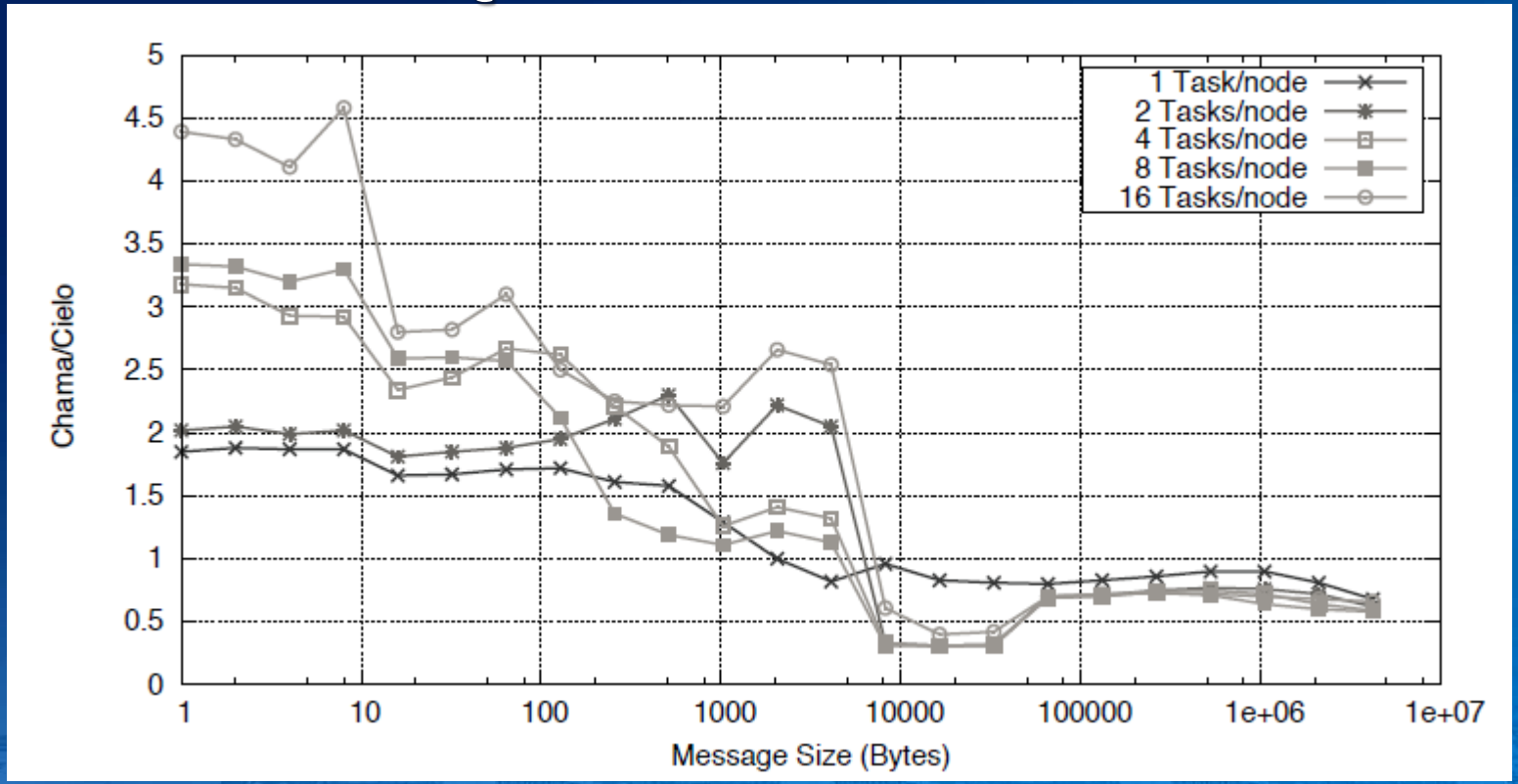
Weak Scaling Zrad3D – 3D Radiation problem's average zone-iteration grind time per machine



Weak Scaling Zrad3D – 3D Radiation problem's average zone-iteration grind time per machine



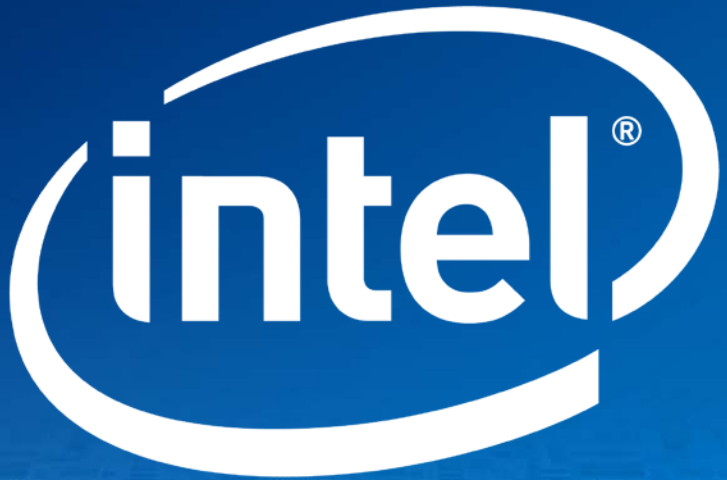
QDR-40/SeaStar2+



From: Unprecedented Scalability and Performance of the New NNSA Tri-Lab Linux Capacity Cluster 2
M. Rajan, D.W. Doerfler, P.T. Lin, S.D. Hammond, R.F. Barrett, and C.T. Vaughan

Summary

- Integration of memory and network into processor will help keep us on the path to Exascale
- Energy is the overwhelming challenge. We need a balanced attack that optimizes energy under real user conditions
- B:F and memory/core while they have their place, they can also result in impediments to progress
- Commodity interconnect can deliver scalability through improvements in Bandwidth, Latency and message rates


intel

Legal Information

Today's presentations contain forward-looking statements. All statements made that are not historical facts are subject to a number of risks and uncertainties, and actual results may differ materially. Please refer to our most recent Earnings Release and our most recent Form 10-Q or 10-K filing for more information on the risk factors that could cause actual results to differ.

If we use any non-GAAP financial measures during the presentations, you will find on our website, intc.com, the required reconciliation to the most directly comparable GAAP financial measure.

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, reference www.intel.com/software/products.

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.

This slide MUST be used with any slides removed from this presentation

Legal Disclaimers

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. Go to: http://www.intel.com/products/processor_number

Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM). Functionality, performance or other benefits will vary depending on hardware and software configurations. Software applications may not be compatible with all operating systems. Consult your PC manufacturer. For more information, visit http://www.intel.com/go/virtualization

No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology (Intel® TXT) requires a computer system with Intel® Virtualization Technology, an Intel TXT-enabled processor, chipset, BIOS, Authenticated Code Modules and an Intel TXT-compatible measured launched environment (MLE). Intel TXT also requires the system to contain a TPM v1.s. For more information, visit http://www.intel.com/technology/security

Requires a system with Intel® Turbo Boost Technology. Intel Turbo Boost Technology and Intel Turbo Boost Technology 2.0 are only available on select Intel® processors. Consult your PC manufacturer. Performance varies depending on hardware, software, and system configuration. For more information, visit http://www.intel.com/go/turbo

Intel® AES-NI requires a computer system with an AES-NI enabled processor, as well as non-Intel software to execute the instructions in the correct sequence. AES-NI is available on select Intel® processors. For availability, consult your reseller or system manufacturer. For more information, see http://software.intel.com/en-us/articles/intel-advanced-encryption-standard-instructions-aes-ni/