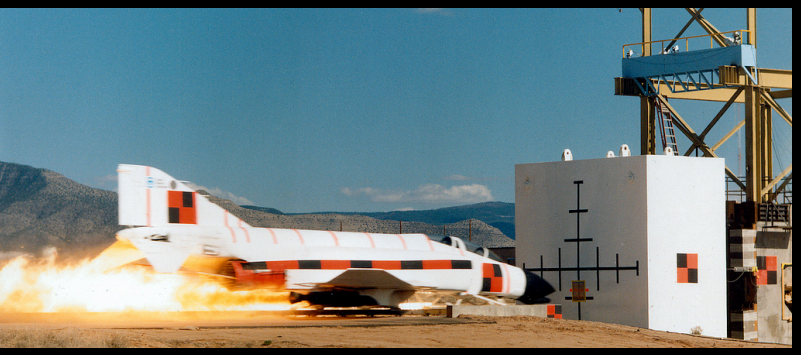
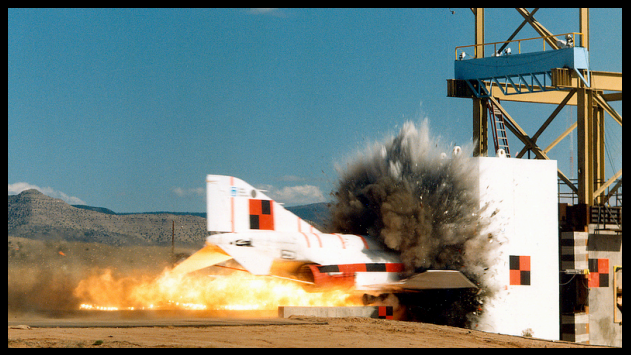
Exceptional service in the national interest

Sandia National Laboratories


Oh, $#*@! Exascale!

The effect of emerging architectures on scientific discovery

Ultrascale Visualization Workshop, November 12, 2012

Kenneth Moreland, Sandia National Laboratories

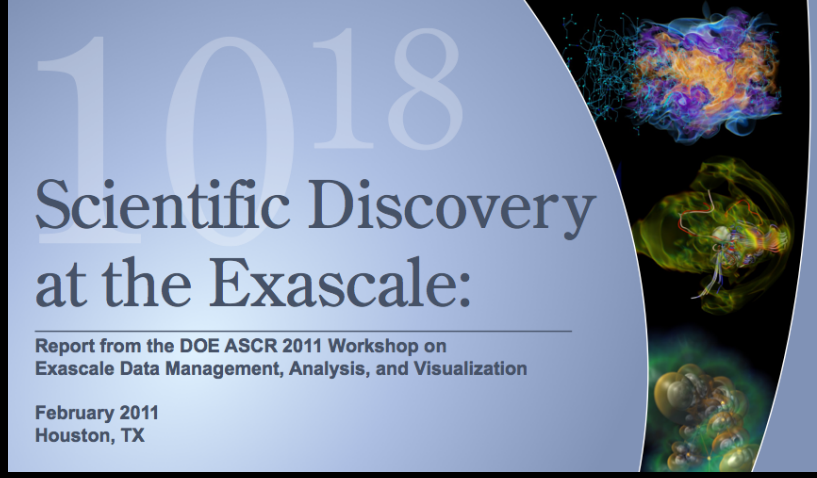
DOE SDAV

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

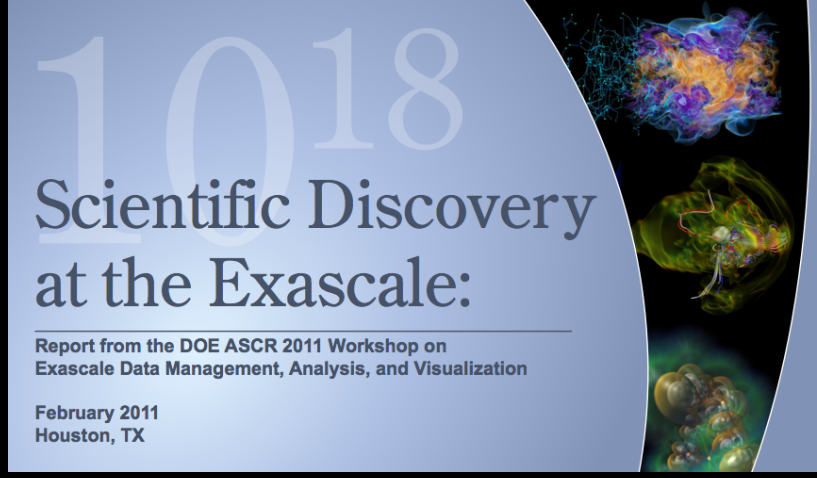
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND 2012-9387P

Slide of Doom



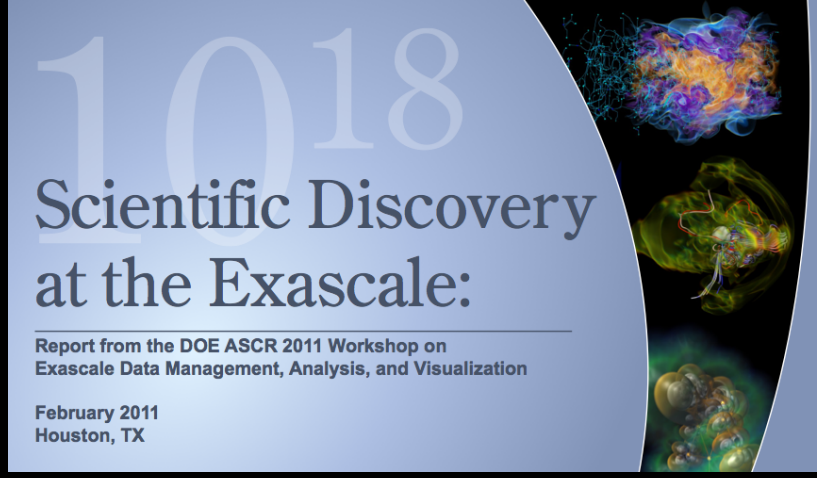
System Parameter	2011	"2018"		Factor Change
System Peak	2 PetaFLOPS	1 ExaFLOP		500
Power	6 MW	≤ 20 MW		3
System Memory	0.3 PB	32 – 64 PB		100 – 200
Total Concurrency	225K	1B × 10	1B × 100	40,000 – 400,000
Node Performance	125 GF	1 TF	10 TF	8 – 80
Node Concurrency	12	1,000	10,000	83 – 830
Network BW	1.5 KB/s	100 GB/s	1000 GB/s	66 – 660
System Size (nodes)	18,700	1,000,000	100,000	50 – 500
I/O Capacity	15 PB	300 – 1000 PB		20 – 67
I/O BW	0.2 TB/s	20 – 60 TB/s		10 – 30

Slide of Doom



System Parameter	2011	"2018"		Factor Change
System Peak	2 PetaFLOPS	1 ExaFLOP		500
Power	6 MW	≤ 20 MW		3
System Memory	0.3 PB	32 – 64 PB		100 – 200
Total Concurrency	225K	1B × 10	1B × 100	40,000 – 400,000
Node Performance	125 GF	1 TF	10 TF	8 – 80
Node Concurrency	12	1,000	10,000	83 – 830
Network BW	1.5 KB/s	100 GB/s	1000 GB/s	66 – 660
System Size (nodes)	18,700	1,000,000	100,000	50 – 500
I/O Capacity	15 PB	300 – 1000 PB		20 – 67
I/O BW	0.2 TB/s	20 – 60 TB/s		10 – 30

Slide of Doom



System Parameter	2011	"2018"		Factor Change
System Peak	2 PetaFLOPS	1 ExaFLOP		500
Power	6 MW	≤ 20 MW		3
System Memory	0.3 PB	32 – 64 PB		100 – 200
Total Concurrency	225K	1B × 10	1B × 100	40,000 – 400,000
Node Performance	125 GF	1 TF	10 TF	8 – 80
Node Concurrency	12	1,000	10,000	83 – 830
Network BW	1.5 KB/s	100 GB/s	1000 GB/s	66 – 660
System Size (nodes)	18,700	1,000,000	100,000	50 – 500
I/O Capacity	15 PB	300 – 1000 PB		20 – 67
I/O BW	0.2 TB/s	20 – 60 TB/s		10 – 30

Exascale Projection

	Jaguar – XT5	Exascale*	Increase
Memory	300 Terabytes	32 – 64 Petabytes	100 – 200×
Concurrency	224,256 way	10 – 100 billion way	Up to 400,000×

MPI Only?

- Vis object code + state: 20MB
- On Jaguar: 20MB × 200,000 processes = 4TB
- On Exascale: 20MB × 100 billion processes = 2EB !

*Source: Scientific Discovery at the Exascale, Ahern, Shoshani, Ma, et al.

Exascale Projection

	Jaguar – XT5	Exascale*	Increase
Memory	300 Terabytes	32 – 64 Petabytes	100 – 200×
Concurrency	224,256 way	10 – 100 billion way	Up to 400,000×

Visualization pipeline too heavyweight?

- On Jaguar: 1 trillion cells → 5 million cells/thread
- On Exascale: 100 trillion cells → 1000 cells/thread

*Source: Scientific Discovery at the Exascale, Ahern, Shoshani, Ma, et al.

Exascale Projection

	Jaguar – XT5	Exascale*	Increase
Memory	300 Terabytes	32 – 64 Petabytes	100 – 200×
Concurrency	224,256 way	10 – 100 billion way	Up to 400,000×

Overhead of ghost/halo cells?

- On Jaguar: 1 trillion cells → 5 million cells/thread
  - Partition into ~171^3 blocks
  - $6 \times 171^2 \approx 175K$ ghost/block → 35 billion ghost total
  - Ghost cells **~3.5% size of original data**
- On Exascale: 100 trillion cells → 1000 cells/thread
  - Partition into 10^3 blocks
  - $6 \times 10^2 \approx 600$ ghost/block → 60 trillion ghost total
  - Ghost cells **60% size of original data**

*Source: Scientific Discovery at the Exascale, Ahern, Shoshani, Ma, et al.

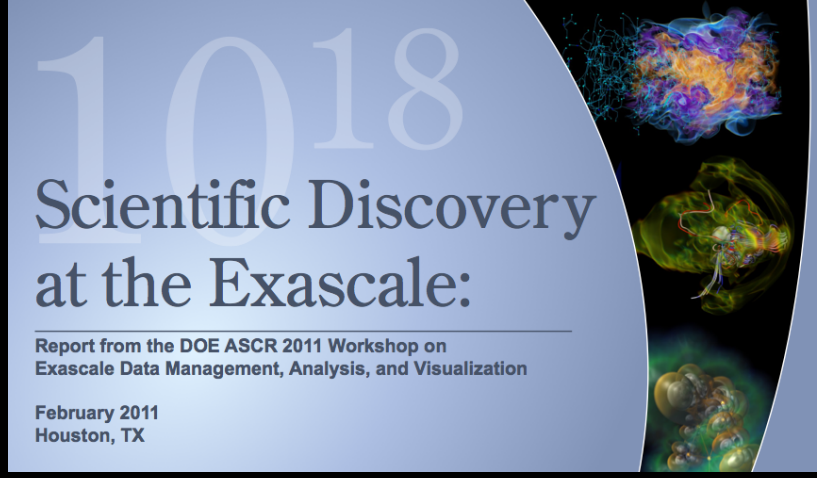
Exascale Programming Challenges

- At some point, domain decomposition fails
  - Too many halo cells, too much communication
- Possible new architectures and programming models
  - GPU accelerators hate decomposition
- Threaded (OpenMP) programming is easier than distributed (MPI) programming.

LIES!!!

  - Threading needs careful planning for memory affinity (inherent in distributed)
  - Sharing memory locations invites read/write collisions (explicit in distributed)
  - PGAS will save us? I'm skeptical.
- Best practice approach: Encapsulated Multithreaded Operations
  - Multiple DOE projects underway: Dax (ASCR), PISTON (ASC), EAVL (LDRD)
  - If successful, minimal impact on applications
  - Might be some changes in scope of what can be done

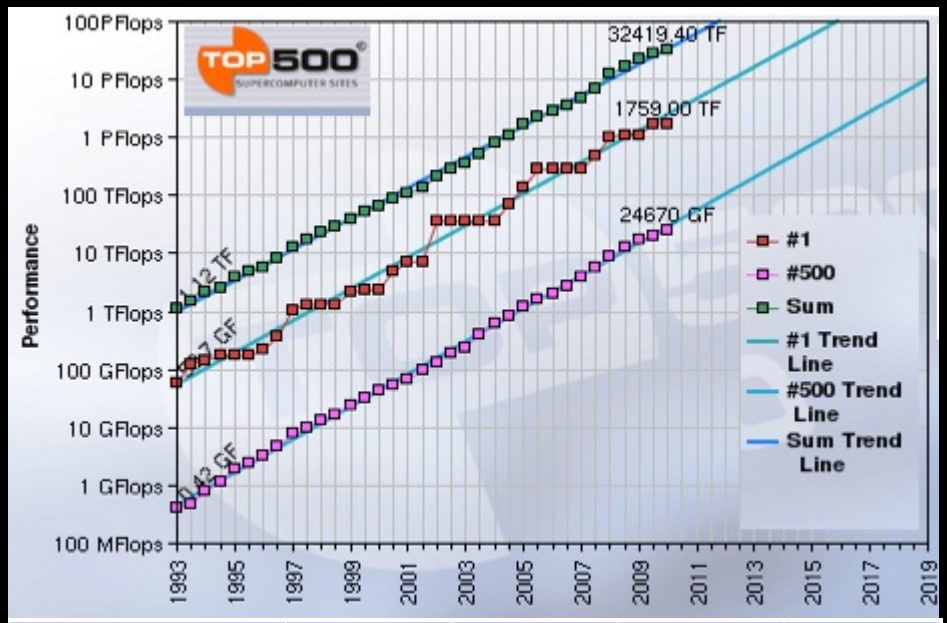
Slide of Doom



System Parameter	2011	"2018"		Factor Change
System Peak	2 PetaFLOPS	1 ExaFLOP		500
Power	6 MW	≤ 20 MW		3
System Memory	0.3 PB	32 – 64 PB		100 – 200
Total Concurrency	225K	1B × 10	1B × 100	40,000 – 400,000
Node Performance	125 GF	1 TF	10 TF	8 – 80
Node Concurrency	12	1,000	10,000	83 – 830
Network BW	1.5 KB/s	100 GB/s	1000 GB/s	66 – 660
System Size (nodes)	18,700	1,000,000	100,000	50 – 500
I/O Capacity	15 PB	300 – 1000 PB		20 – 67
I/O BW	0.2 TB/s	20 – 60 TB/s		10 – 30

Extreme scale computing

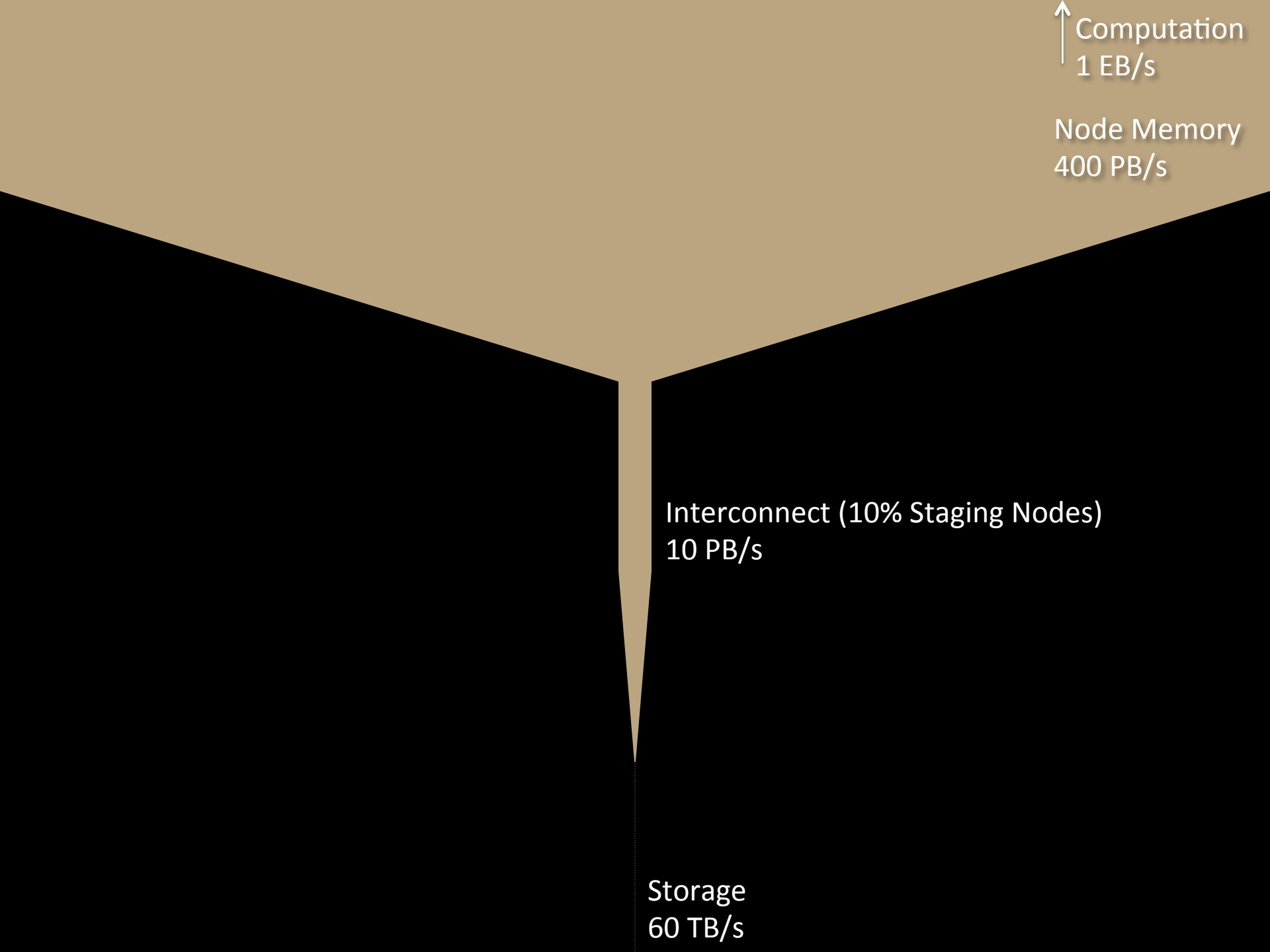
- Trends
  - More FLOPS
  - More concurrency
  - Comparatively less storage, I/O bandwidth
- ASCI purple (49 TB/140 GB/s) – JaguarPF (300 TB/ 200 GB/s)
  - Most people get < 5 GB/sec at scale

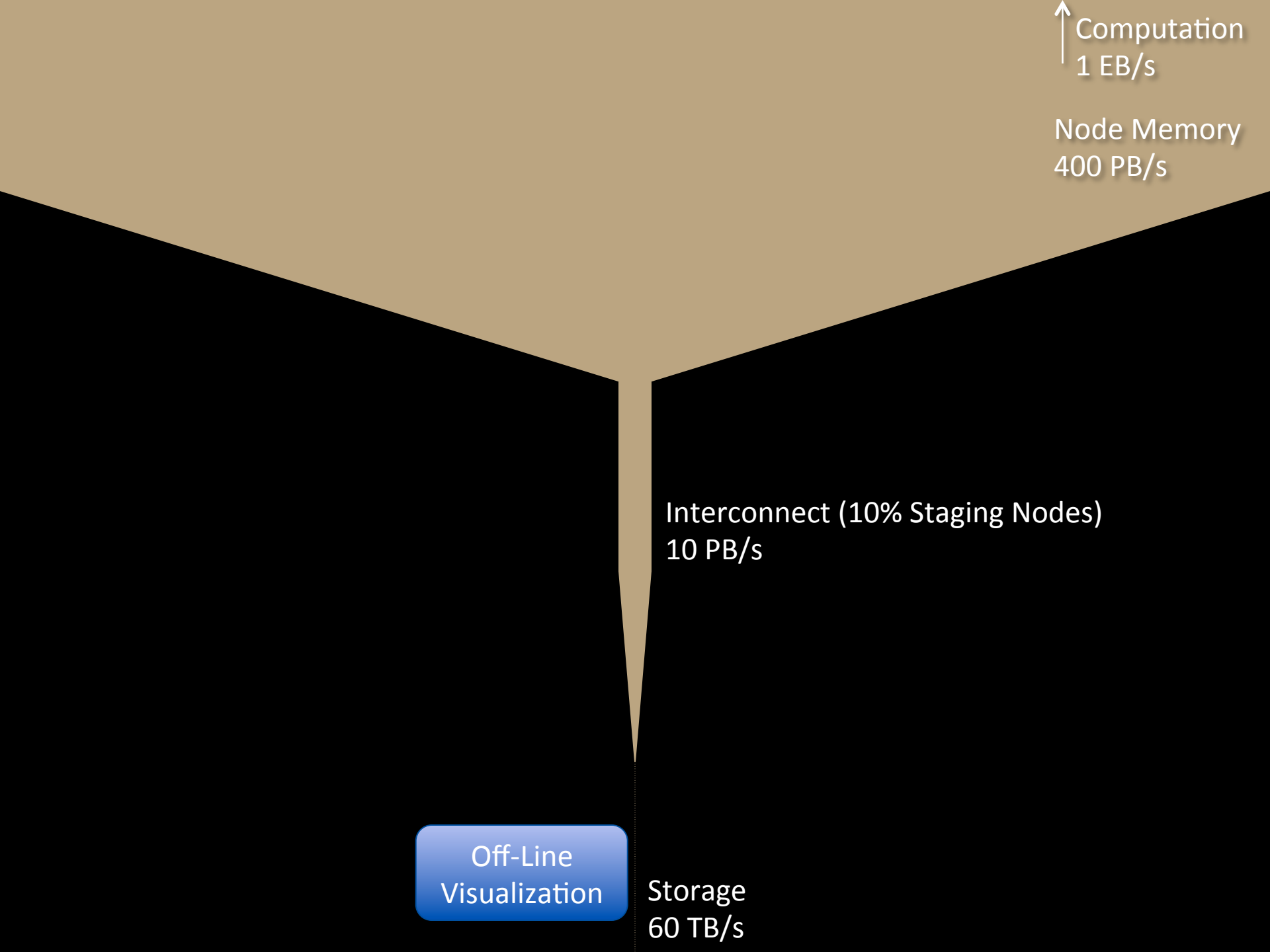


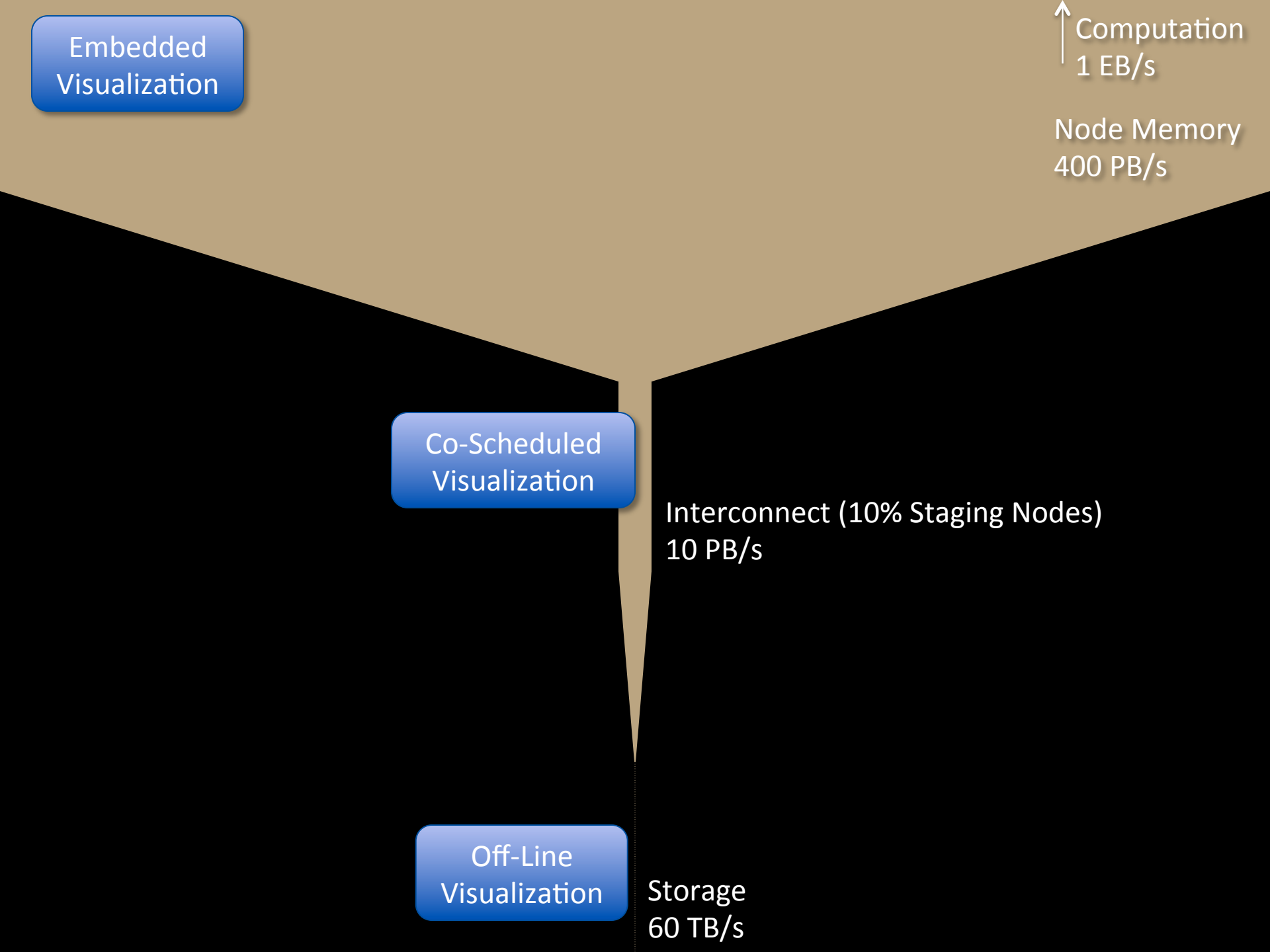
Systems	2009	2011	2015	2018
System Peak Flops/s	2 Peta	20 Peta	100-200 Peta	1 Exa
System Memory	0.3 PB	1 PB	5 PB	10 PB
Node Performance	125 GF	200 GF	400 GF	1-10 TF
Node Memory BW	25 GB/s	40 GB/s	100 GB/s	200-400 GB/s
Node Concurrency	12	32	O(100)	O(1000)
Interconnect BW	1.5 GB/s	10 GB/s	25 GB/s	50 GB/s
System Size (Nodes)	18,700	100,000	500,000	O(Million)
Total Concurrency	225,000	3 Million	50 Million	O(Billion)
Storage	15 PB	30 PB	150 PB	300 PB
I/O	0.2 TB/s	2 TB/s	10 TB/s	20 TB/s
MTTI	Days	Days	Days	O(1Day)
Power	6 MW	~10 MW	~10 MW	~20 MW

ILLUSTRATION: A. TOVEY

From J. Dongarra, "Impact of Architecture and Technology for Extreme Scale on Software and Algorithm Design," Cross-cutting Technologies for Computing at the Exascale, February 2-5, 2010.


Computation
1 EB/s
Node Memory
400 PB/s
Interconnect (10% Staging Nodes)
10 PB/s
Storage
60 TB/s


Computation
1 EB/s
Node Memory
400 PB/s
Interconnect (10% Staging Nodes)
10 PB/s
Off-Line
Visualization
Storage
60 TB/s


Embedded Visualization
Computation
1 EB/s
Node Memory
400 PB/s
Co-Scheduled Visualization
Interconnect (10% Staging Nodes)
10 PB/s
Off-Line Visualization
Storage
60 TB/s

Space of Solutions

	Capability	Coupling	Footprint	Transfer	Interactive
Tightly Integrated	Low	Tight	**Low**	**None**	No
Embedded	**High**	Tight	High	Possible memcpy	No
Hybrid	**High**	Tight	Medium	Subset Hi Speed Transfer	**Yes**
Co-Scheduled	**High**	**Loose**	~5% Extra Nodes	Hi Speed Transfer	**Yes**
Off-Line	**High**	**Loose**	**None**	Slow Persistent Storage Cost	**Yes**

Other Exascale Challenges

- Resilience
  - Mean time to failure
  - Robustness for in situ
  - Soft errors
  - Uncertainty
- Compression/extraction
  - Feature characterization
  - Lossy/lossless compression
- Provenance
  - Capture functions, algorithms, parameters
- Uncertainty Quantification

Acknowledgements

- Funding support
  - The DOE Office of Science, Advanced Scientific Computing Research, under award number 10-014707, program manager Lucy Nowell
  - The Director, Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. 12-015215, through the Scientific Discovery through Advanced Computing (SciDAC) Institute of Scalable Data Management, Analysis and Visualization
- Collaborators
  - Sandia National Laboratories: Kenneth Moreland, Nathan Fabian, Ron Oldfield
  - Los Alamos National Laboratory: James Ahrens, Jonathan Woodring
  - Oak Ridge National Laboratory: Scott Klasky, Norbert Podhorszki
  - Argonne National Laboratory: Venkatram Vishwanath, Mark Hereld, Michael E. Papka
  - Kitware, Inc.: Berk Geveci, Utkarsh Ayachit, Andrew C. Bauer, Pat Marion, Sebastien Jourdain, David DeMarle, David Thompson
  - University of Colorado at Boulder: Michel Rasquin, Kenneth E. Jansen
  - Rutgers University: Ciprian Docan, Manish Parashar