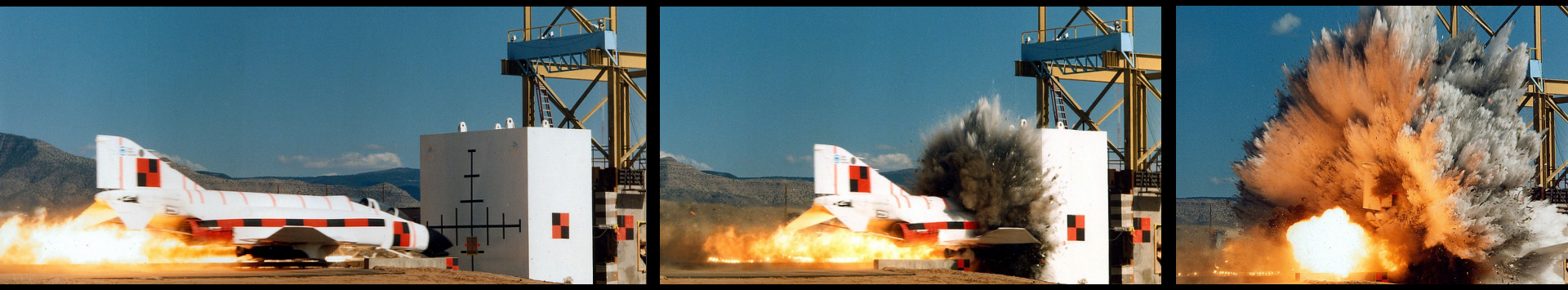
Exceptional service in the national interest

Sandia National Laboratories



Oh, $#*@! Exascale!

The effect of emerging architectures on data analysis software

SOS 17 Panel, March 27, 2013

Kenneth Moreland, Sandia National Laboratories

SDAV

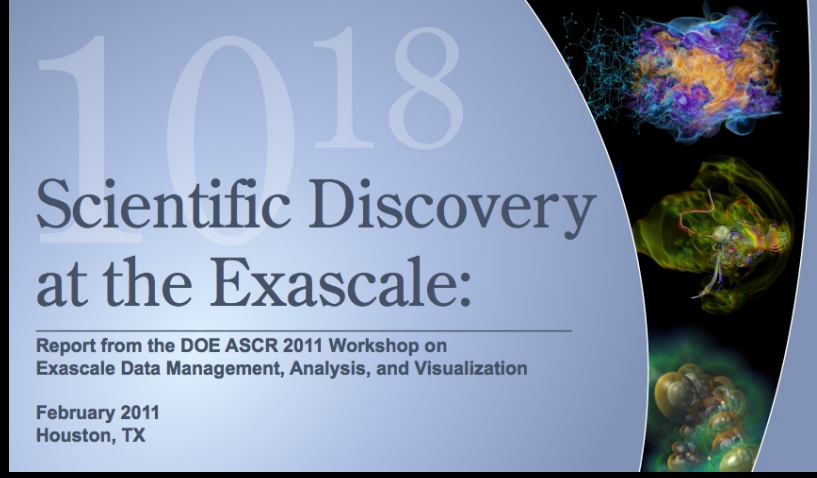
U.S. DEPARTMENT OF ENERGY

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.

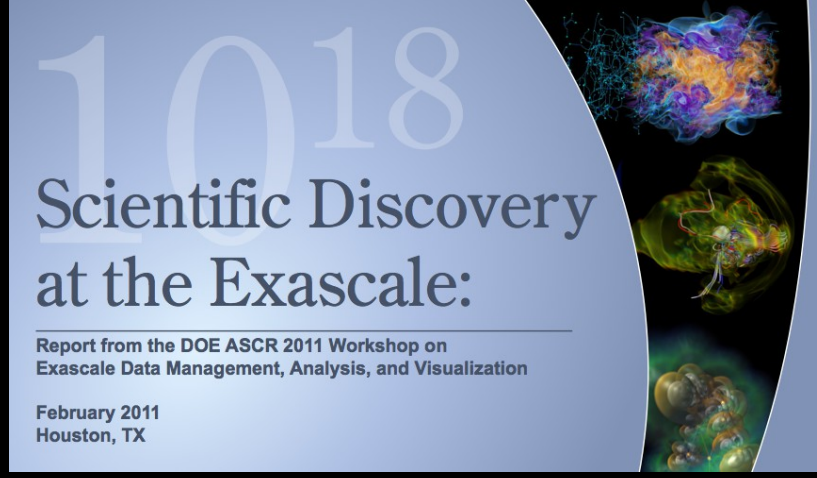
SAND 2013-2003P

Slide of Doom



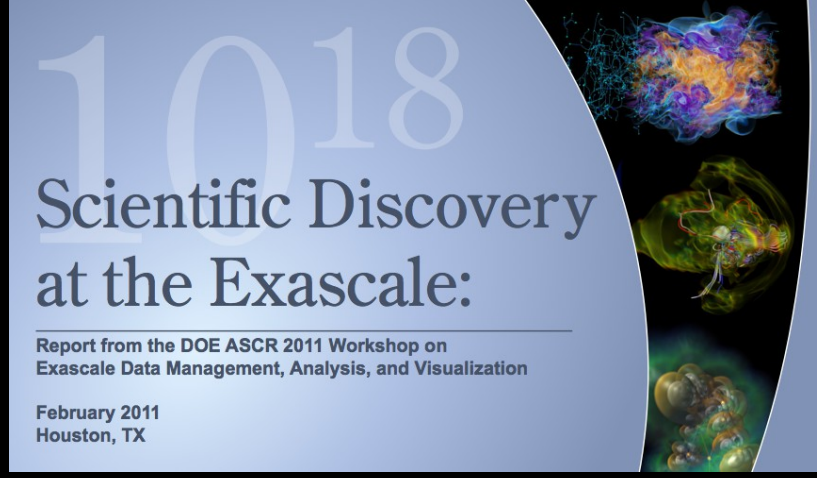
System Parameter	2011	"2018"		Factor Change
System Peak	2 PetaFLOPS	1 ExaFLOP		500
Power	6 MW	≤ 20 MW		3
System Memory	0.3 PB	32 – 64 PB		100 – 200
Total Concurrency	225K	1B × 10	1B × 100	40,000 – 400,000
Node Performance	125 GF	1 TF	10 TF	8 – 80
Node Concurrency	12	1,000	10,000	83 – 830
Network BW	1.5 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

What is Extreme Scale?

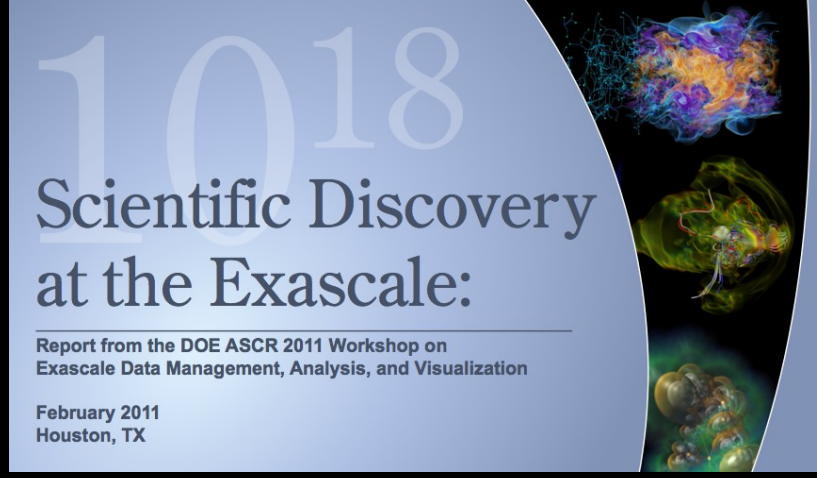
	Jaguar – XT5	Titan – XK7	Exascale*
Cores	224,256	299,008 and 18,688 gpu	1 billion
Concurrency	224,256 way	70 – 500 millon way	10 – 100 billion way
Memory	300 Terabytes	700 Terabytes	128 Petabytes

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

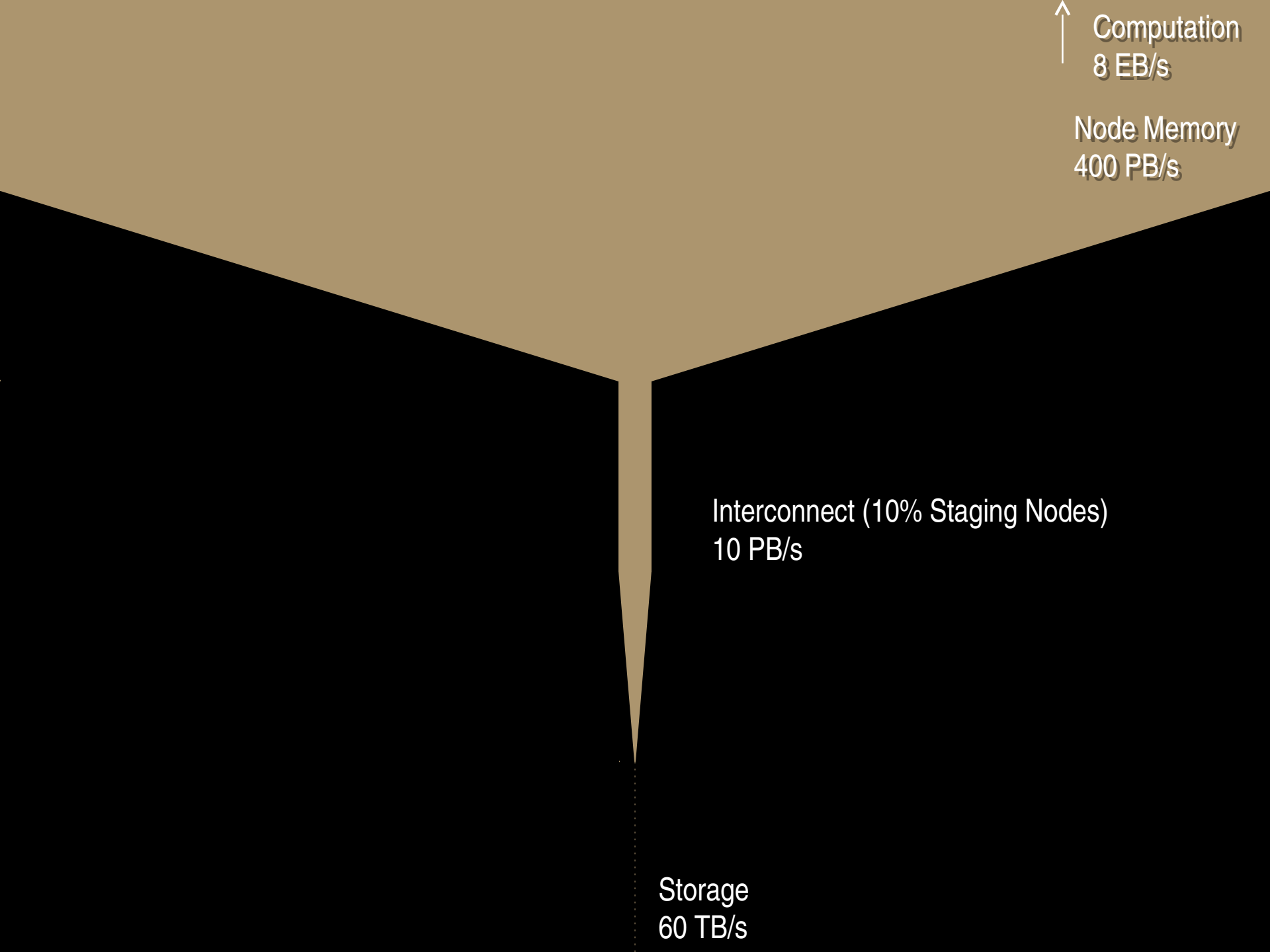
Exascale Programming Challenges

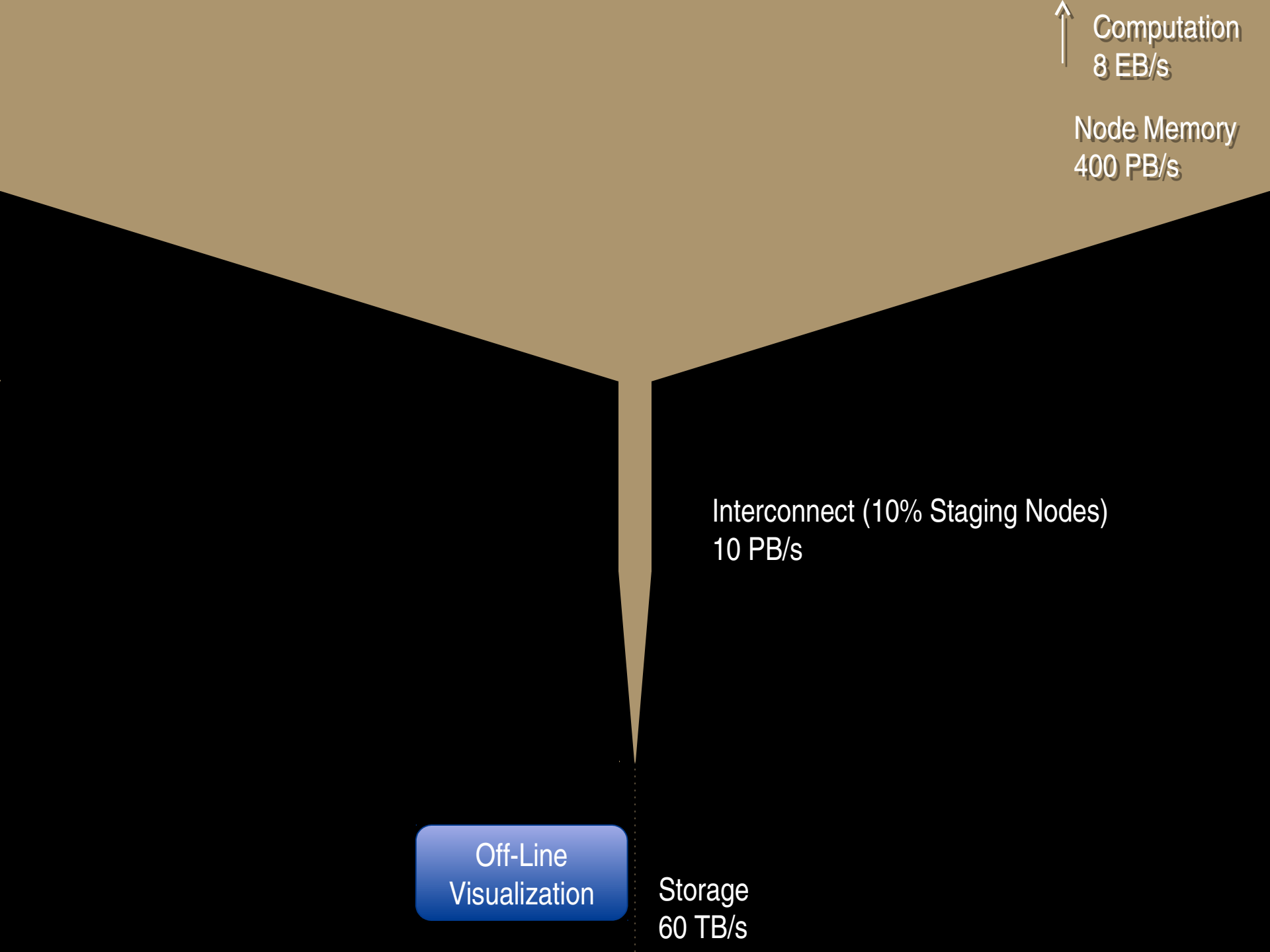
- At some point, domain decomposition fails
  - Too many halo cells, too much communication
- Possible new architectures and programming models
  - GPU accelerators hate decomposition
- Threaded (OpenMP) programming is easier than distributed (MPI) programming.
  - Threading needs careful planning for 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: Parallel Functor application (Map, Visitor)
  - Multiple DOE projects underway: Dax (ASCR), PISTON (ASC), EAVL (LDRD)
  - If successful, minimal impact on applications

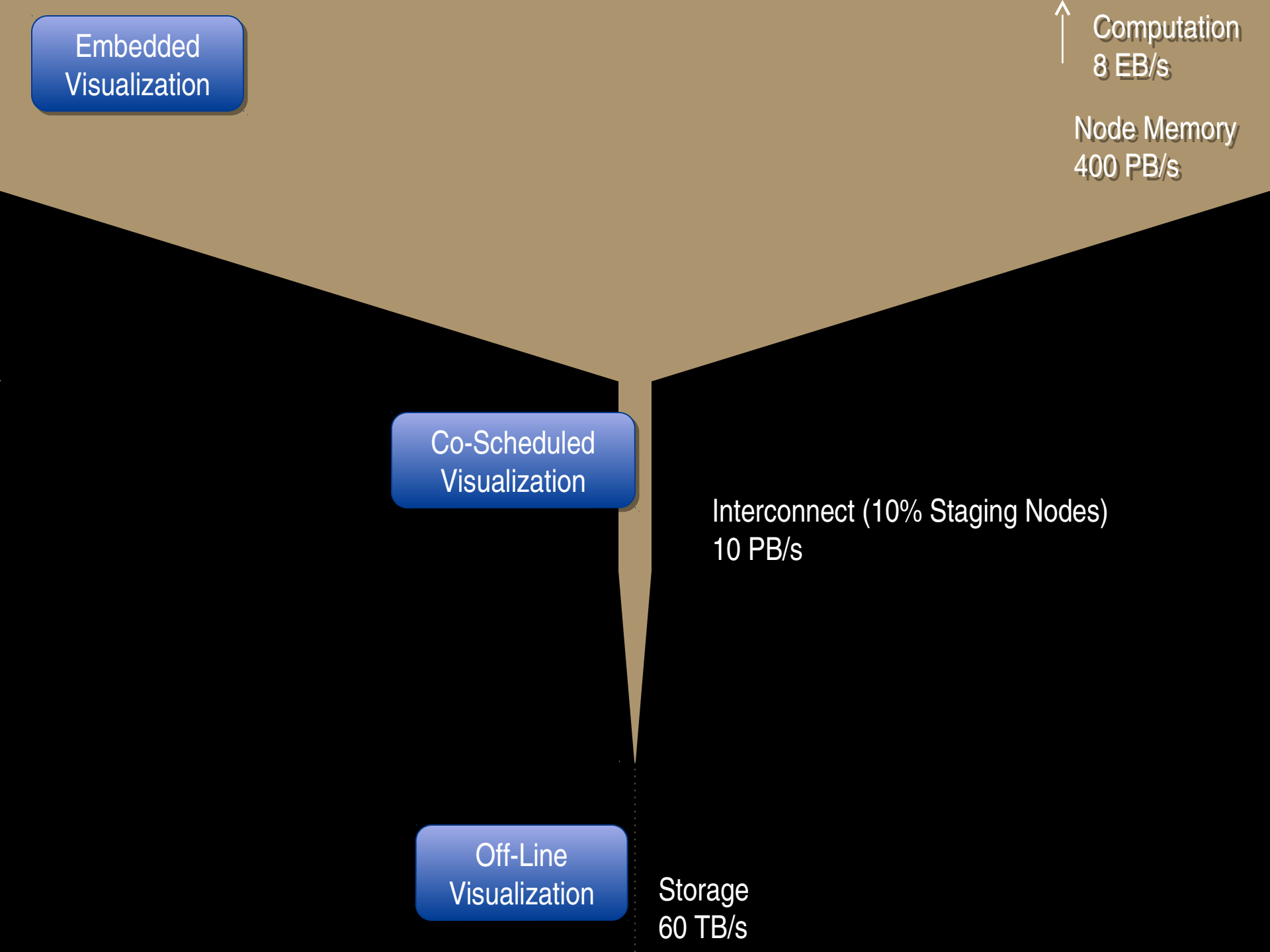
Slide of Doom



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


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


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


Embedded Visualization
Computation
8 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 Challenges (Cop Out Slide)

- 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