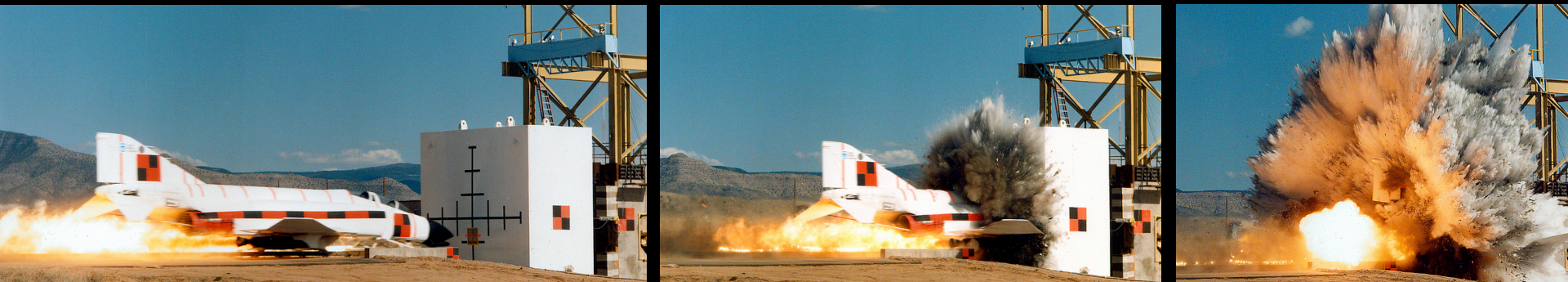


Exceptional service in the national interest



Oh, \$#* @! Exascale!

The effect of emerging architectures on data science

CScADS Panel, August 2, 2012

Kenneth Moreland, Sandia National Laboratories



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 2012-5994P

Slide of Doom

1018

Scientific Discovery at the Exascale:

Report from the DOE ASCR 2011 Workshop on
Exascale Data Management, Analysis, and Visualization

February 2011
Houston, TX



| 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 \times 10 | 1B \times 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

1018

Scientific Discovery at the Exascale:

Report from the DOE ASCR 2011 Workshop on
Exascale Data Management, Analysis, and Visualization

February 2011
Houston, TX



| 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

1018

Scientific Discovery at the Exascale:

Report from the DOE ASCR 2011 Workshop on
Exascale Data Management, Analysis, and Visualization

February 2011
Houston, TX



| 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 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: Parallel Functor application (Map, Visitor)
 - 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

1018

Scientific Discovery at the Exascale:

Report from the DOE ASCR 2011 Workshop on
Exascale Data Management, Analysis, and Visualization

February 2011
Houston, TX



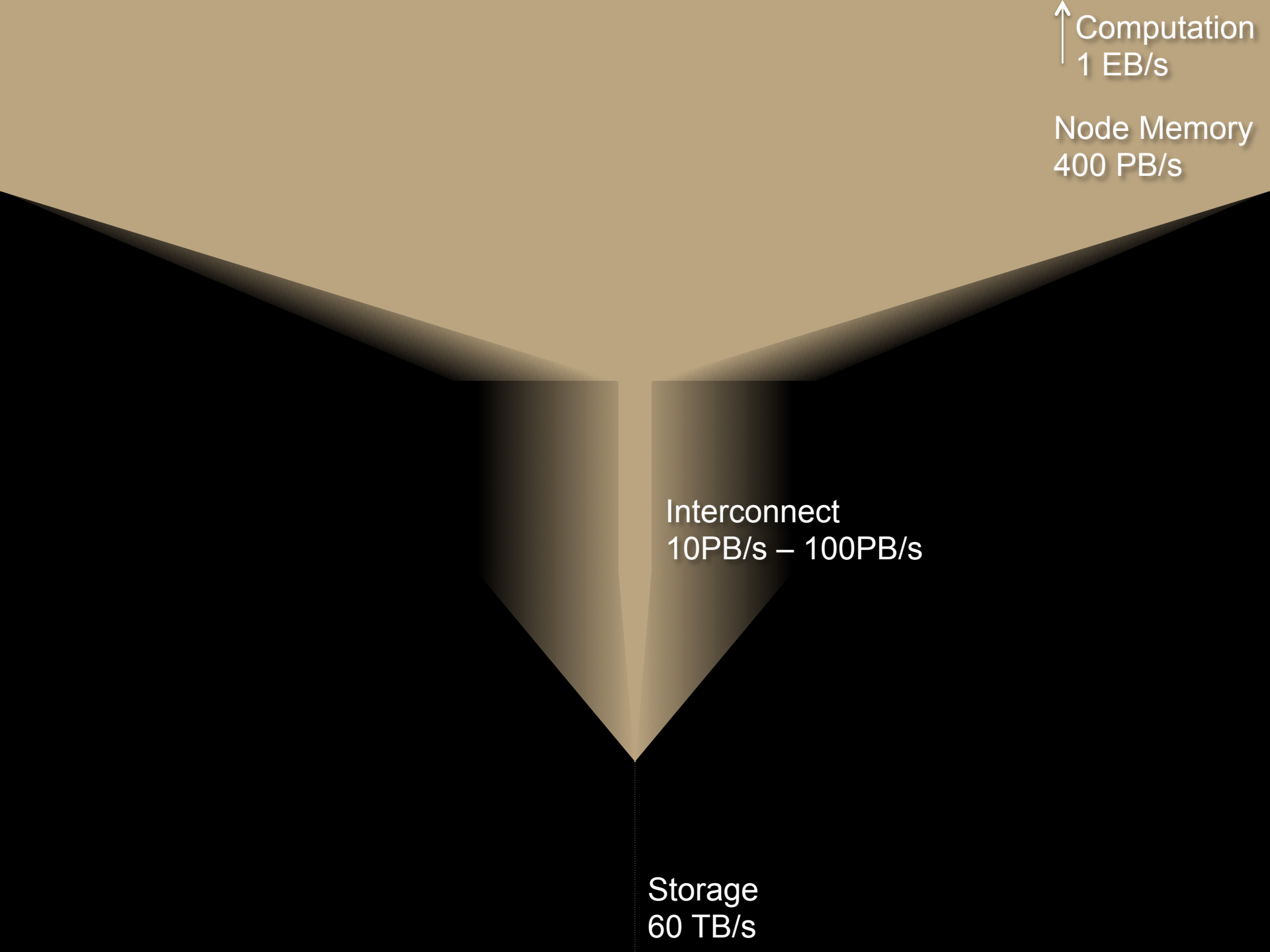
| System Parameter | 2011 | "2018" | | Factor Change |
|---------------------|-------------|---------------|-----------|------------------|
| System Peak | 2 PetaFLOPS | 1 ExaFLOP | | 500 |
| Power | 6 MW | ≤ 20 MW | | 3 |
| System Memory | 0.3 PB | 32 – 64 PB | | 100 – 200 |
| Total Concurrency | 225K | 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
1 EB/s

Node Memory
400 PB/s

Interconnect
10PB/s – 100PB/s

Storage
60 TB/s

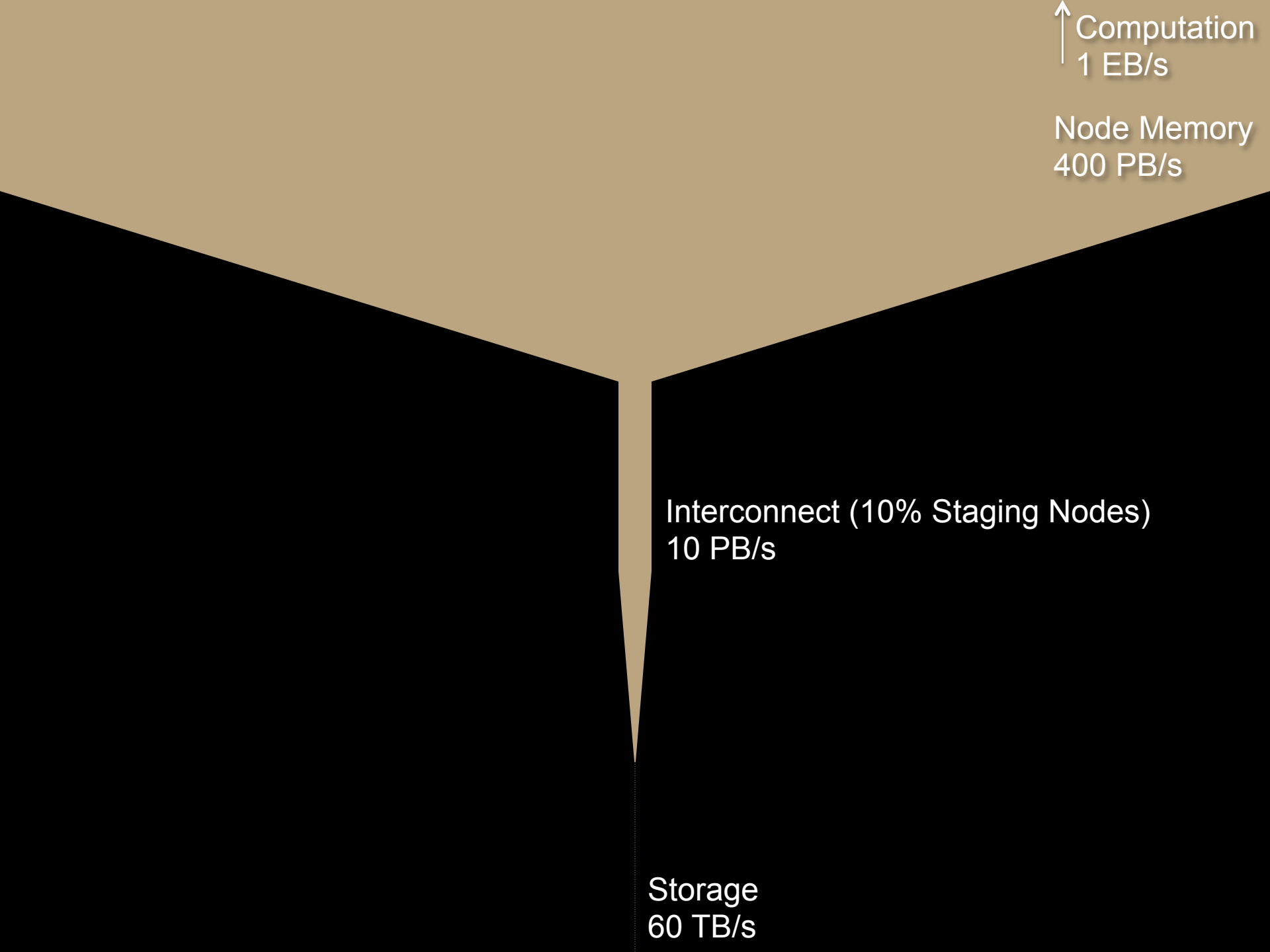


↑ Computation
1 EB/s

Node Memory
400 PB/s

Interconnect (10% Staging Nodes)
10 PB/s

Storage
60 TB/s



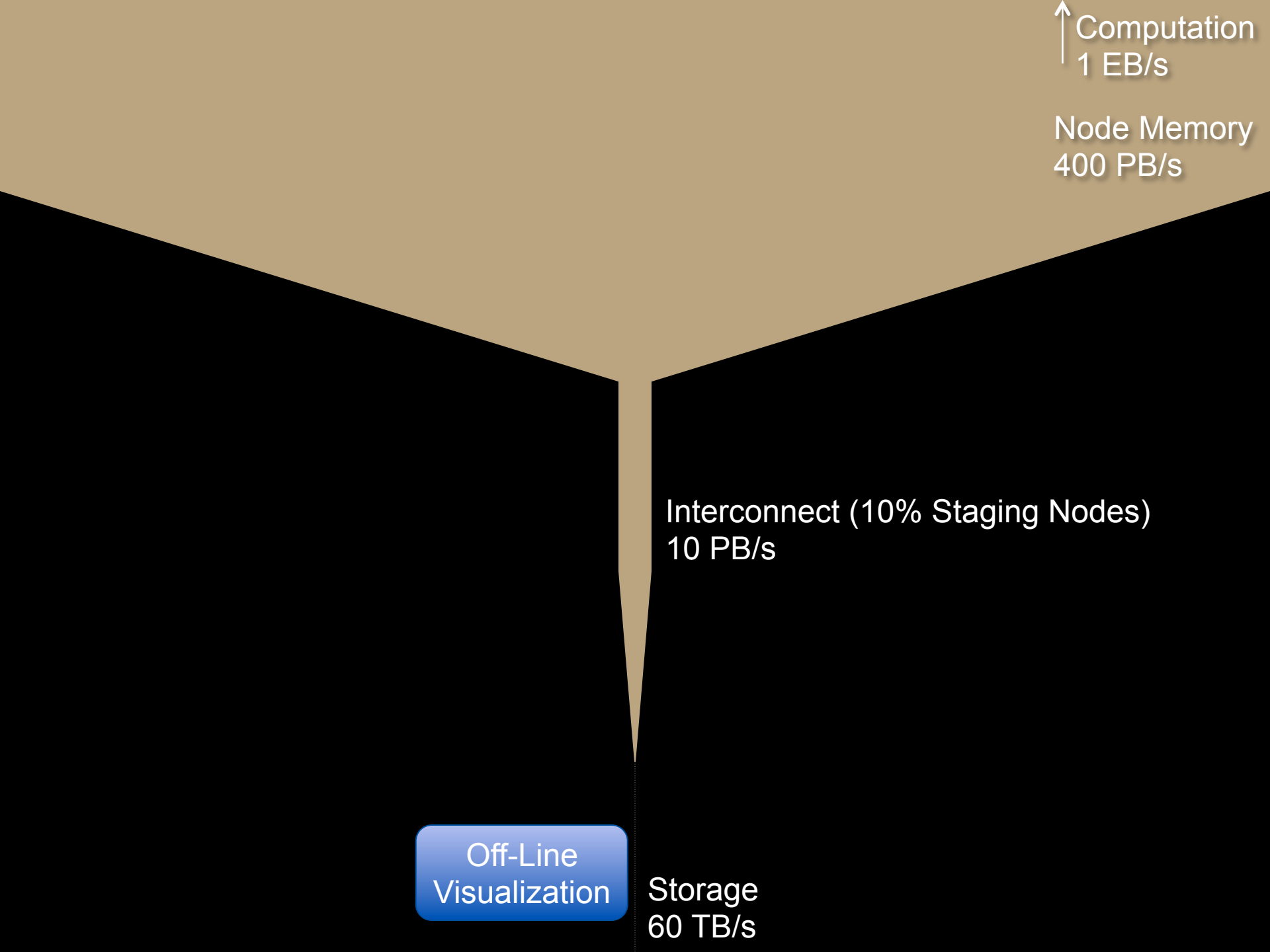
↑ Computation
1 EB/s

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 |