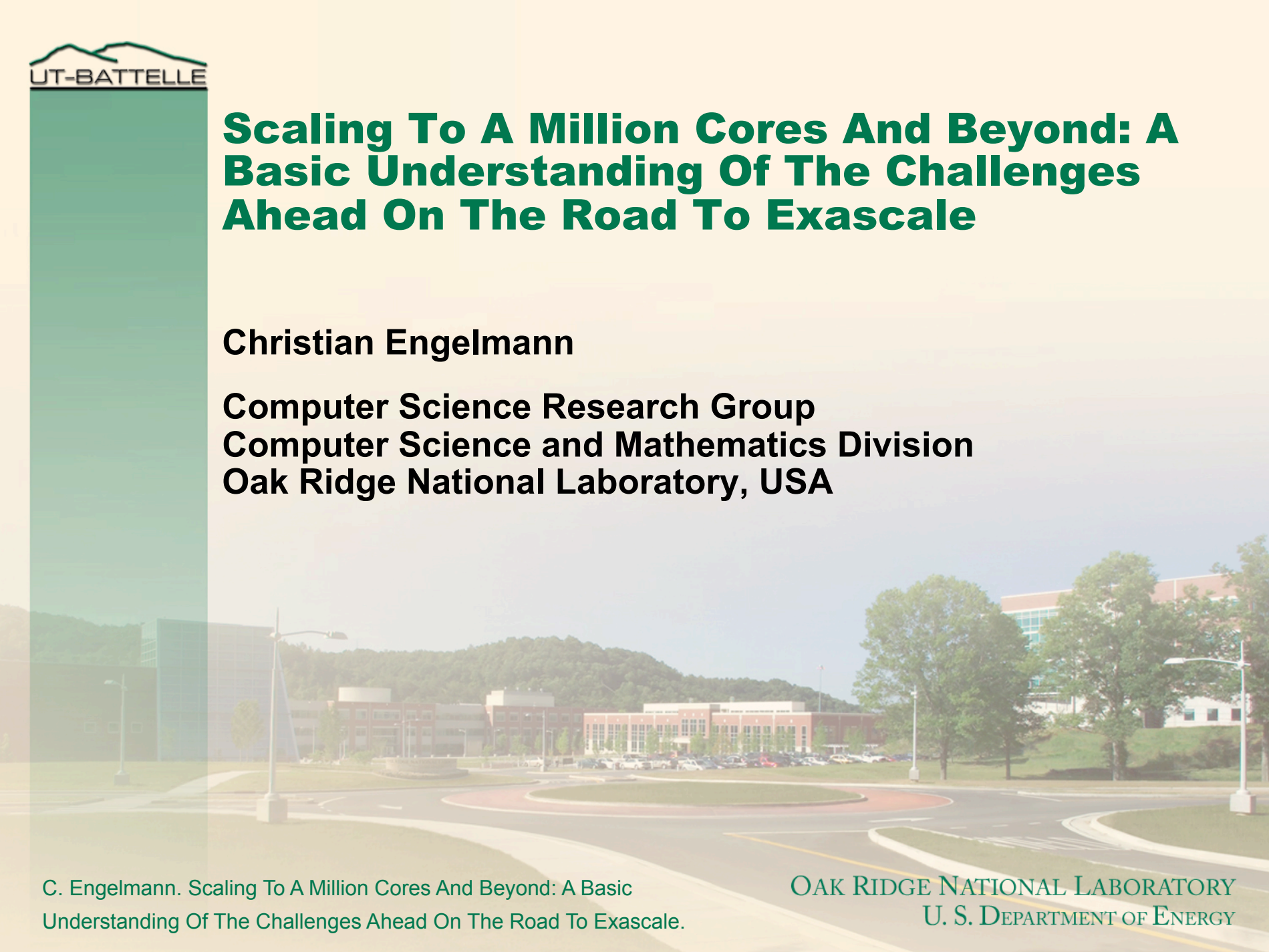
UT-BATTELLE

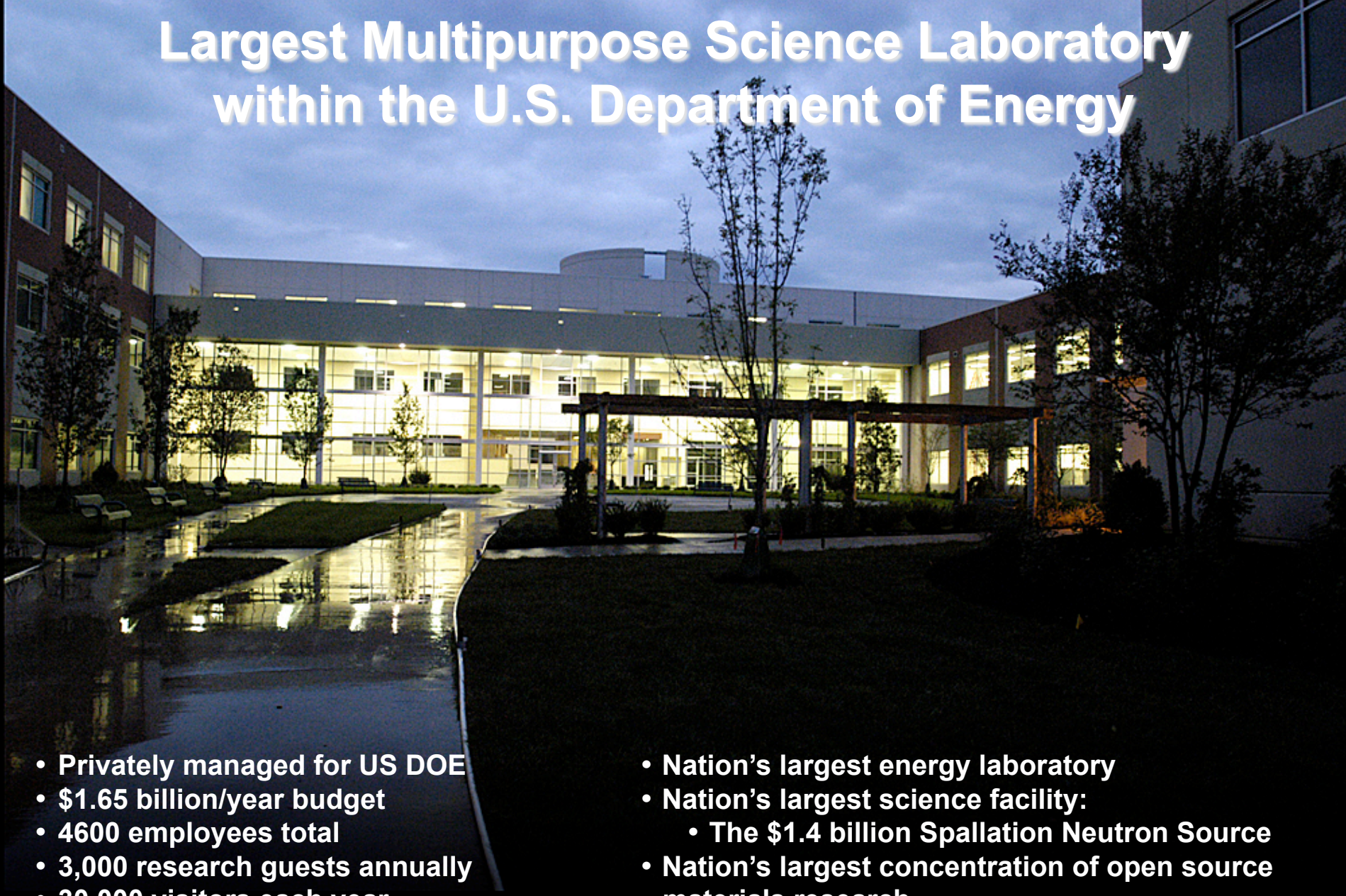
Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale

Christian Engelmann

Computer Science Research Group
Computer Science and Mathematics Division
Oak Ridge National Laboratory, USA



Largest Multipurpose Science Laboratory within the U.S. Department of Energy



- Privately managed for US DOE
- $1.65 billion/year budget
- 4600 employees total
- 3,000 research guests annually
- 30,000 visitors each year
- Total land area 58mi^2 (150km^2)

- Nation's largest energy laboratory
- Nation's largest science facility:
  - The $1.4 billion Spallation Neutron Source
- Nation's largest concentration of open source materials research
- Nation's largest open scientific computing facility

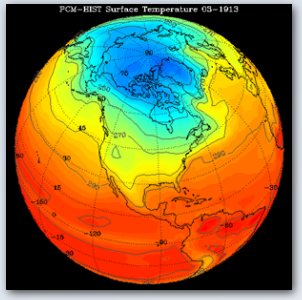
ORNL East Campus: Site of World Leading Computing and Computational Sciences

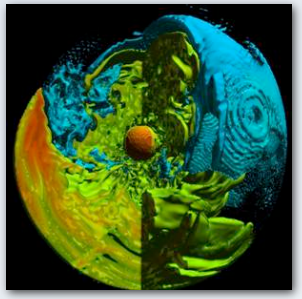
National Center for Computational Sciences

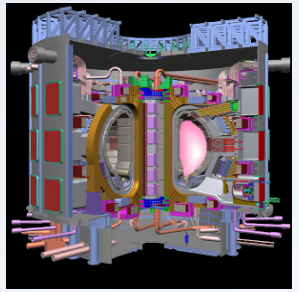
- 40,000 ft^2 (3700 m^2) computer center:
  - 36-in (~1m) raised floor, 18 ft (5.5 m) deck-to-deck
  - 12 MW of power with 4,800 t of redundant cooling
  - High-ceiling area for visualization lab:
    - 35 MPixel PowerWall, Access Grid, etc.

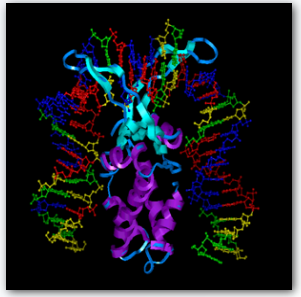
- 4 systems in the Top 500 List of Supercomputer Sites:
  - Jaguar: 3. Cray XT5-HE, MPP with 224,162 cores ➪ 1.759 PFlop/s LINPACK
  - Kraken (UT): 11. Cray XT5-HE, MPP with 112,800 cores ➪ 1.173 PFlop/s LINPACK
  - Gaea C2 (NOAA): 20. Cray XE6, MPP with 77,824 cores ➪ 0.716 PFlop/s LINPACK
  - Gaea (NOAA): 51. Cray XT6-HE, MPP with 30,912 cores ➪ 0.260 PFlop/s LINPACK

At Forefront in Scientific Computing and Simulation









- Leading partnership in developing the National Leadership Computing Facility
  - Leadership-class scientific computing capability
  - 10-20 Pflop/s by 2012 (Titan, CPU/GPGPU, Cray)
- Attacking key computational challenges
  - Climate change
  - Nuclear astrophysics
  - Fusion energy
  - Materials sciences
  - Biology
- Providing access to computational resources through high-speed networking

C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Trends in HPC System Design

- Ongoing trends in HPC system design:
  - Increasing core counts (in total and per processor)
  - Increasing node counts (OS instances)
  - Heterogeneity (CPU+GPGPU at large scale)
- Emerging technology influencing HPC system design:
  - Stacked memory (3D chip layering)
  - Non-volatile memory (SSD and phase change memory)
  - Network-on-chip
- Additional forces influencing HPC system design:
  - Power consumption ceiling (overall and per-chip)
- *How to design HPC systems to fit application needs?*
- *How to design applications to efficiently use HPC systems?*

C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Current-Generation HPC Systems

- Large-scale PFlop/s systems:
  - #1 RIKEN K: 8.162 PFlop/s, 548,352 cores, 93%
  - #2 NSCT Tianhe-1A: 2.566 PFlop/s, 186,368 cores, 55%
  - #3 ORNL Jaguar XT5: 1.759 PFlop/s, 224,162 cores, 75%
  - #4 NSCS Nebulae: 1.271 PFlop/s, 120,640 cores, 43%
  - #5 GSIC Tsubame 2.0: 1.192 PFlop/s, 73,278 cores, 61%
  - #5 LANL Cielo: 1.110 PFlop/s, 142,272 cores, 81%
  - #6 NASA Pleiades: 1.088 PFlop/s, 111,104 cores, 81%
  - #7 LBNL Hopper: 1.054 PFlop/s, 153,408 cores, 82%
- The trend is toward even larger-scale systems
  - End of processor frequency scaling ➔ Node/core scaling
  - New technologies: Chip stacking, NoC, NVRAM, Si photonics

C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Discussed Exascale Road Map

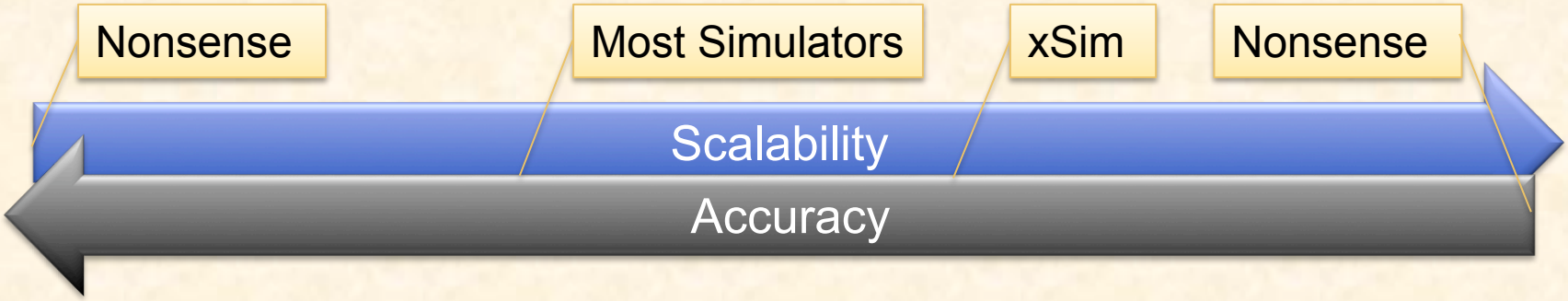
Many design factors are driven by the power ceiling of 20MW

Systems	2009	2012	2016	2020
System peak	2 Peta	20 Peta	100-200 Peta	1 Exa
System memory	0.3 PB	1.6 PB	5 PB	10 PB
Node performance	125 GF	200GF	200-400 GF	1-10TF
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	22 GB/s	25 GB/s	50 GB/s
System size (nodes)	18,700	100,000	500,000	O(million)
Total concurrency	225,000	3,200,000	O(50,000,000)	O(billion)
Storage	15 PB	30 PB	150 PB	300 PB
IO	0.2 TB/s	2 TB/s	10 TB/s	20 TB/s
MTTI	1-4 days	5-19 hours	50-230 min	22-120 min
Power	6 MW	~10MW	~10 MW	~20 MW

C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

xSim: Facilitating HPC Hardware/Software Co-Design Through Simulation

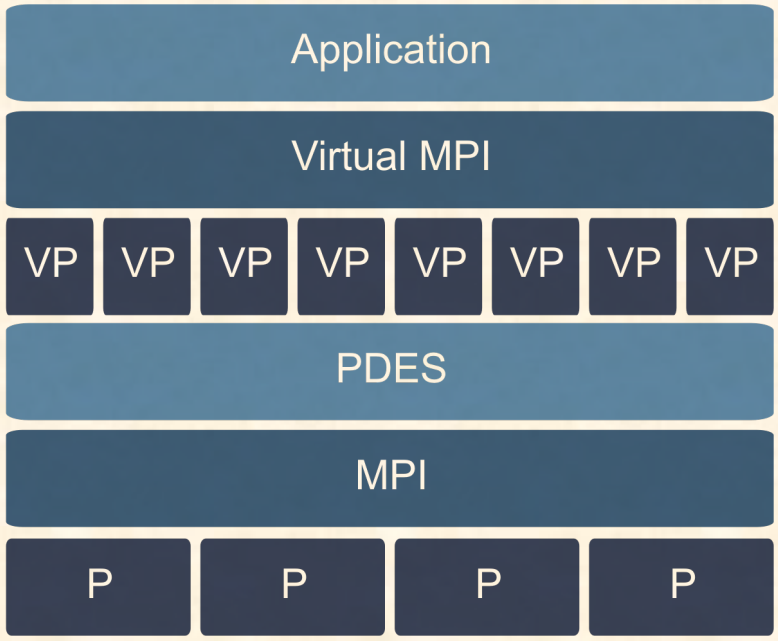
- **Execution of real applications, algorithms or their models atop a simulated HPC environment for:**
  - **Performance evaluation, including identification of resource contention and underutilization issues**
  - **Investigation at extreme scale, beyond the capabilities of existing simulation efforts**
- **xSim: Highly scalable solution that trades off accuracy**



C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

xSim: Technical Approach

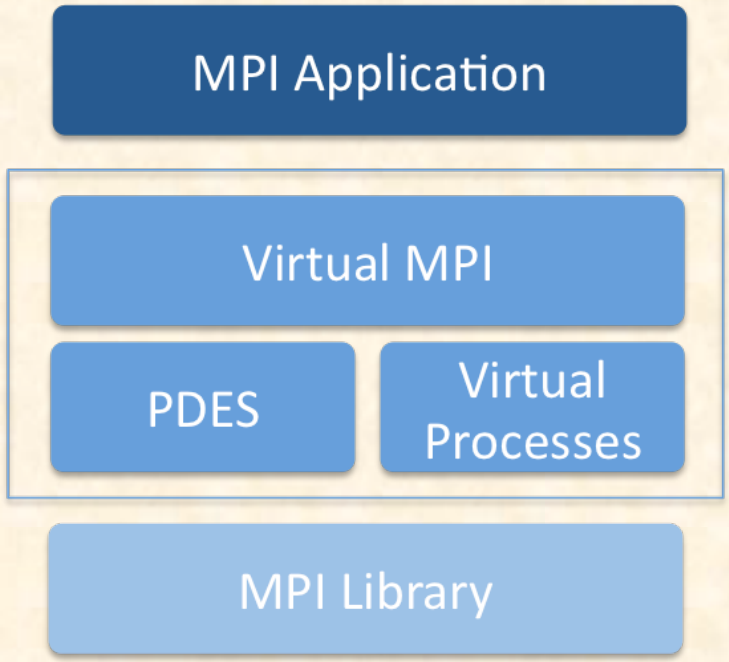
- Combining highly oversub-scribed execution, a virtual MPI, and a time-accurate PDES
- PDES uses the native MPI and simulates virtual processors
- The virtual processors expose a virtual MPI to applications
- Applications run within the context of virtual processors:
  - Global and local virtual time
  - Execution on native processor
  - Local and native MPI communication
  - Processor and network model



C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

xSim: Design

- The simulator is a library
- Utilizes PMPI to intercept MPI calls and to hide the PDES
- Implemented in C with 2 threads per native process
- Support for C and Fortran MPI
- Easy to use:
  - Compile with xSim header
  - Link with the xSim library
  - Execute:




```
mpirun -np <np> <application> -xsim-np <vp>
```

xSim: Implementation

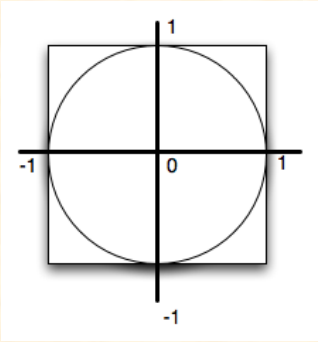
- PDES
  - Virtual time for each VP using actual execution time and processor model
  - Virtual MPI latency/band-width using network model
  - Conservative execution with deadlock detection
- Virtual Processes
  - User-space threads with stack frame and global variables context switch
  - Stack overflow protection
  - Intel 32/64-bit (x86/x86_64)
- Virtual MPI
  - Reimplementation of MPI atop VP P2P messaging
  - MPI groups/communicators
  - MPI collectives/timing
  - Determ. MPI_ANY_SOURCE
- Network model
  - P2P message latency and bandwidth for different network architectures
  - No contention at this time
- Processor model
  - Scaled execution time

Investigating Parallel Algorithm and System Performance Properties at Extreme Scale: A Case Study

- Using a micro application to demonstrate
  - General scaling properties
  - Impact of multi-core scaling
  - Impact of core throttling
- Using xSim for
  - Running micro application at extreme scale
  - Simulating current and future HPC architectures
- Running xSim on
  - 960-core (40-node) cluster
  - AMD, GbE, 2.5TB RAM
- Investigated systems:
  - Perfect network, i.e., 0 latency and ∞ bandwidth
  - Current system similar to Cray XT with 16-core 3D Twisted Torus
  - Current/next-generation system with 128-core 3D Twisted Torus
  - Beyond next-generation system with 1024-core 3D Twisted Torus
  - 128- and 1024-core systems with less powerful cores

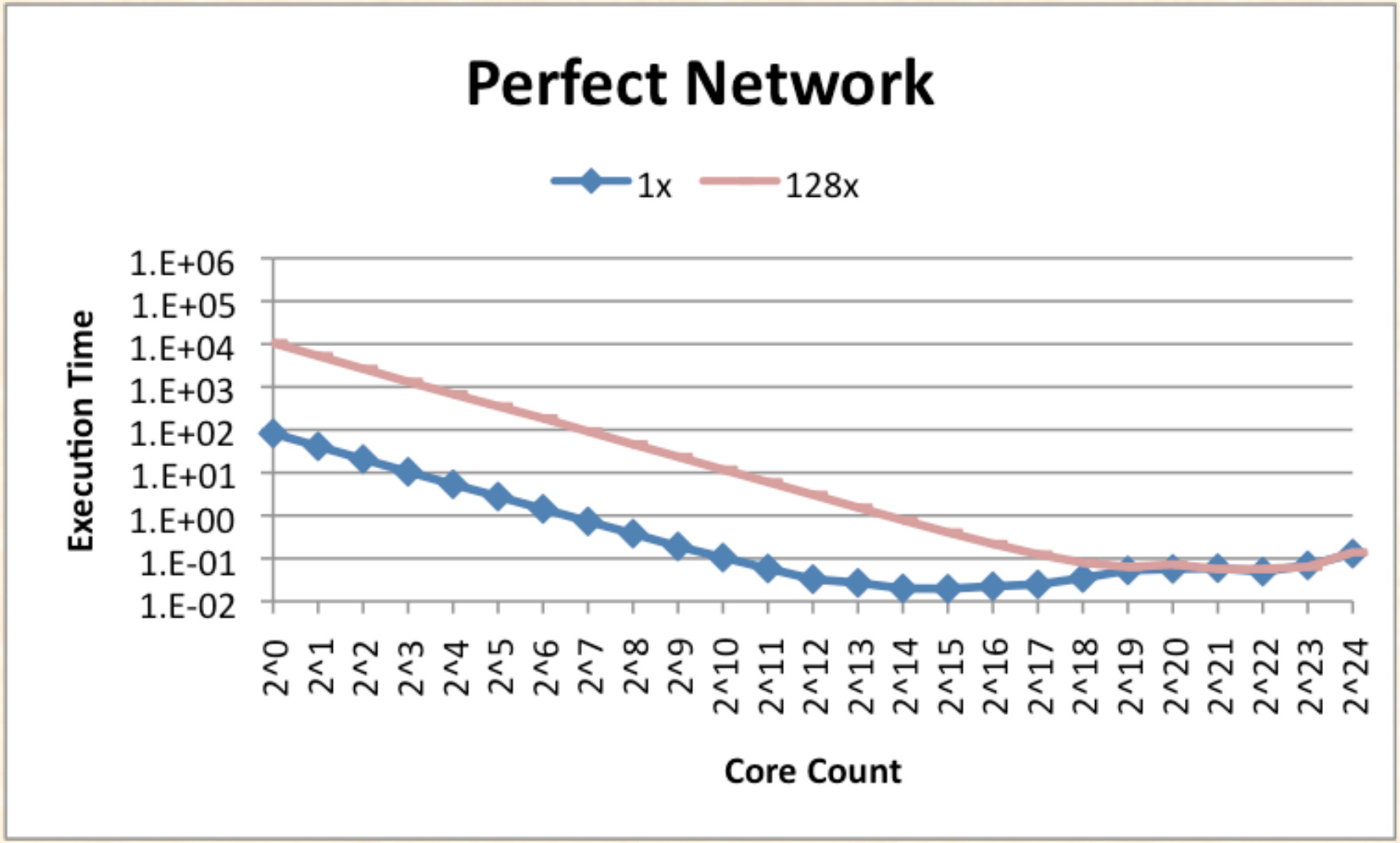
Micro Application: PI Monte Carlo Solver

- Basic Monte Carlo solver
- Estimates the value of PI
- Dartboard approach:
  - Randomly throw darts inside a 2x2 square area
  - Count darts that hit the radius 1 circle
  - PI = 4* hits/total



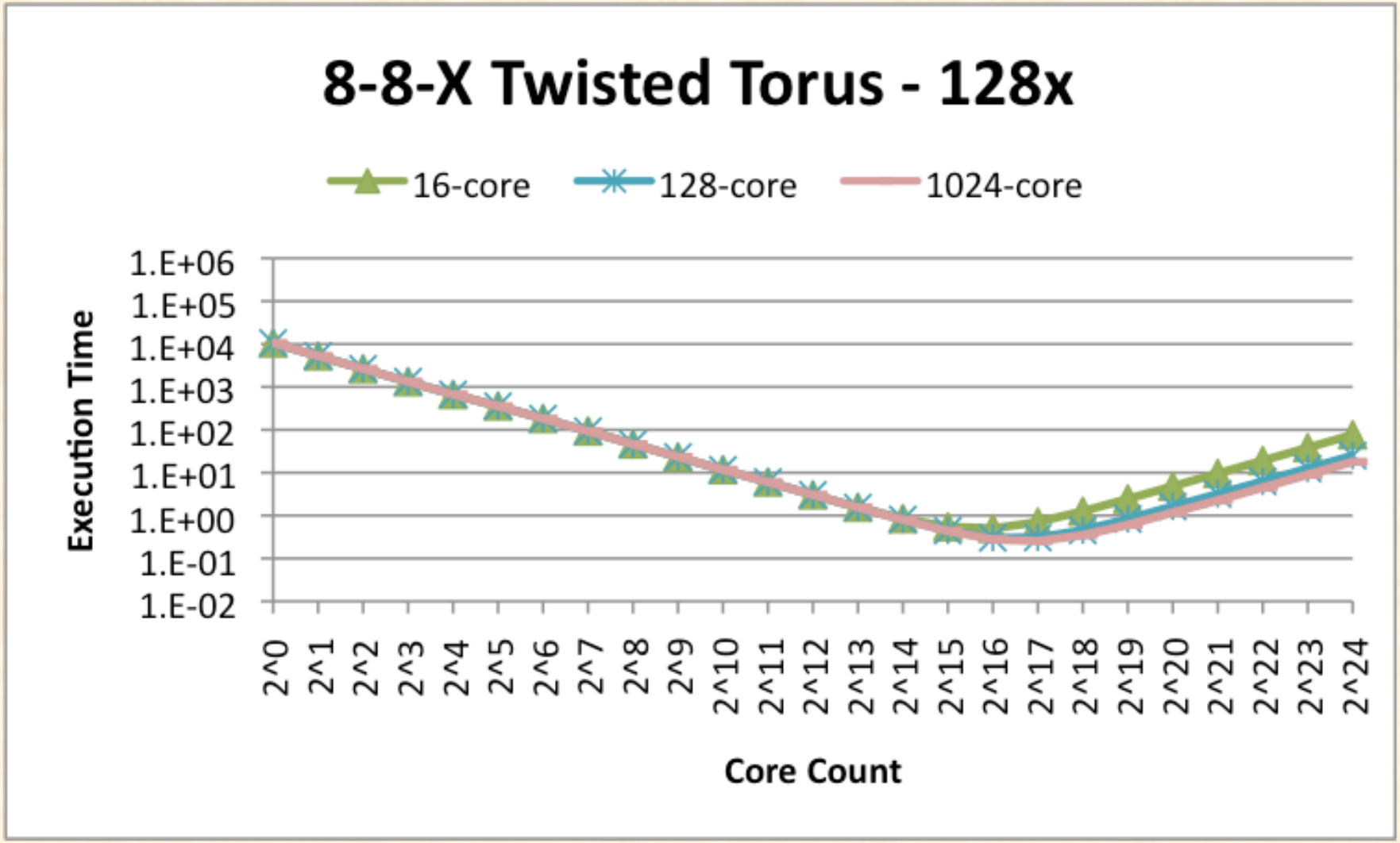
- Nearly embarrassingly parallel solver
  - Parallel generation of random numbers
  - Parallel counting of hits
  - Linear collection of hit count at rank 0 (by design)
  - Final calculation and printout at rank 0
- Scaling expectation
  - Scales linear until sequential part starts to dominate (Amdahl's law)

Scaling with Perfect Network (AMD Opteron CPU, 0 Latency, ∞ Bandwidth)



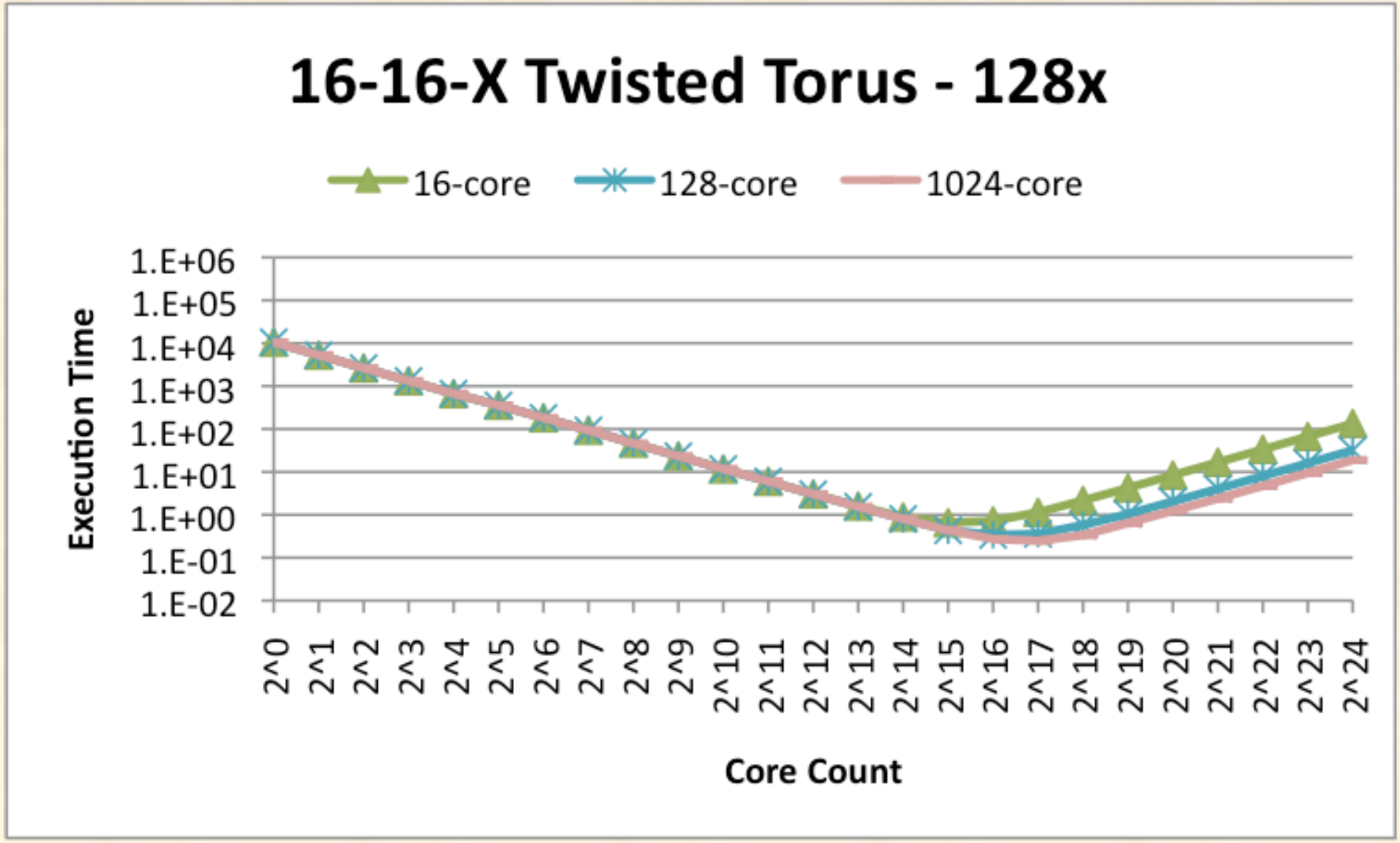
C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Scaling with X-core 3D/8 Twisted Torus (AMD Opteron CPU, Cray XT5 Network)



C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Scaling with X-core 3D/16 Twisted Torus (AMD Opteron CPU, Cray XT5 Network)

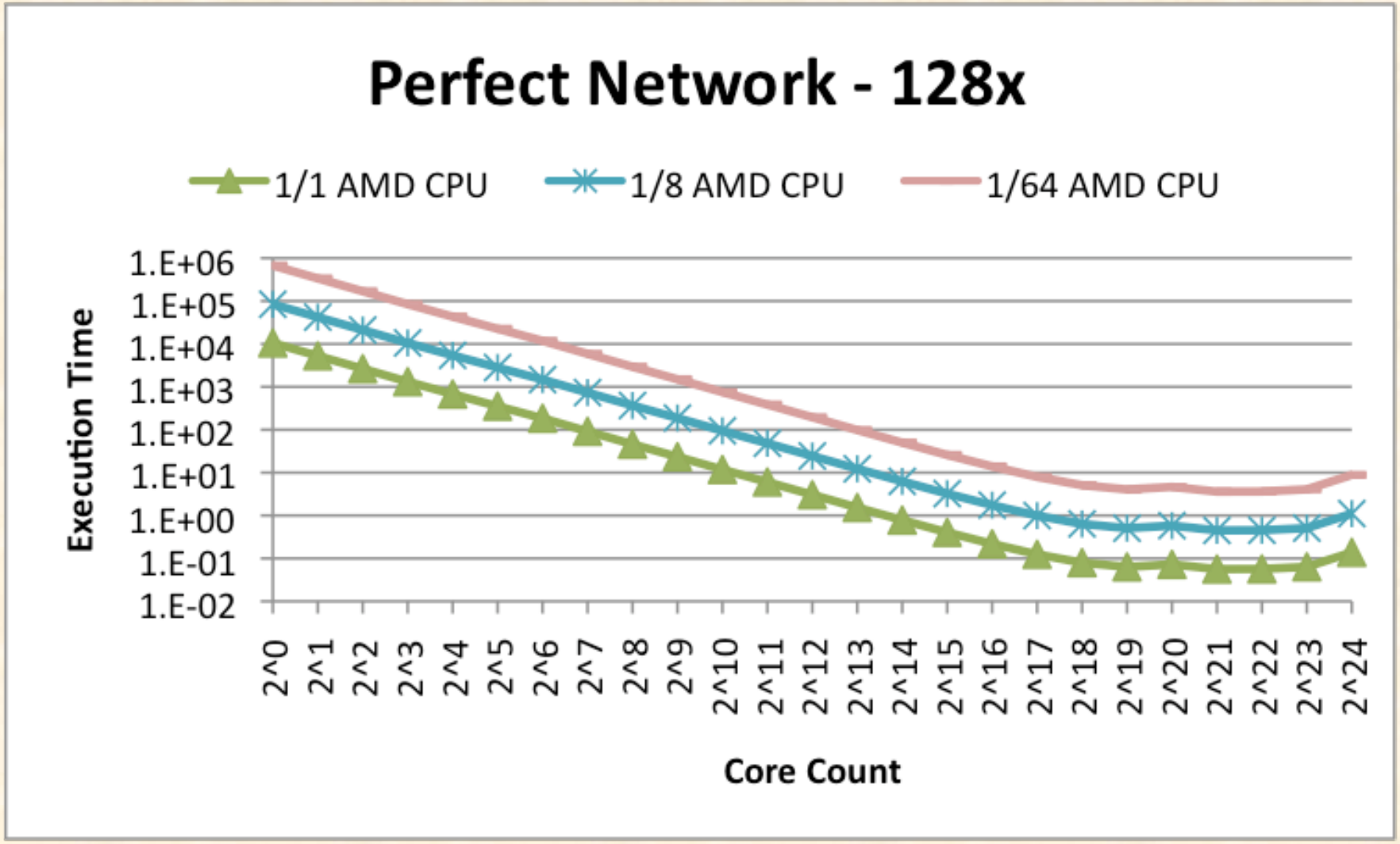


C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Basic Scaling Properties

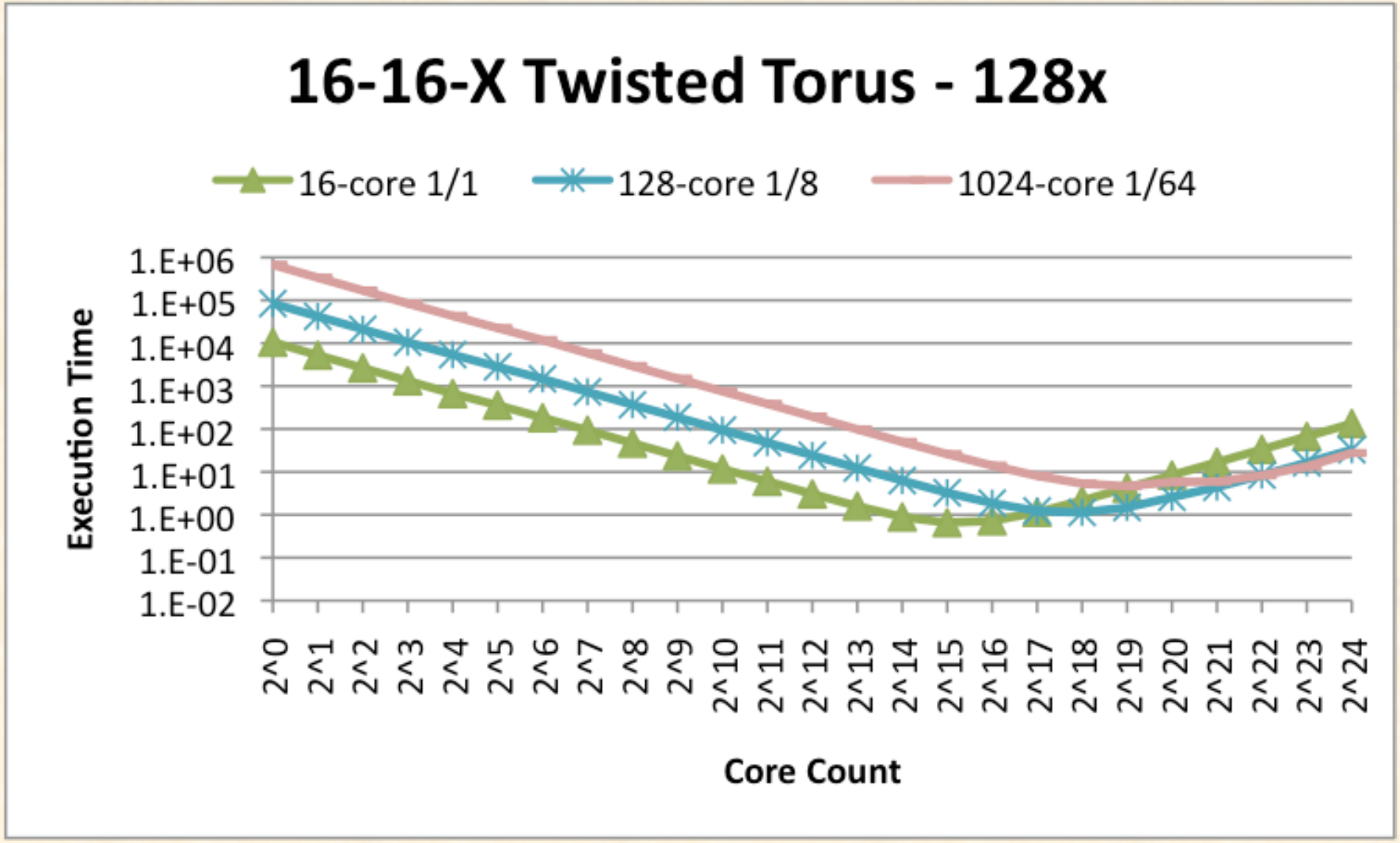
- Perfect network
  - Linear scaling until algorithm runs out of parallel work at 2^18 (262,144) cores
  - Increasing the number of iterations pushes the scalability limit out
- 3-D 8x8x.. Twisted torus
  - Linear scaling until communication starts to dominate at 2^15 (32,768) cores
  - Performance difference between 16- and 128-core
  - Performance difference between 128- and 1024-core minimal
- 3-D 16x16x.. twisted torus
  - Slight performance decrease due to mismatch between communication pattern and twisted torus

Scaling with Perfect Network (X CPU, 0 Latency, ∞ Bandwidth)



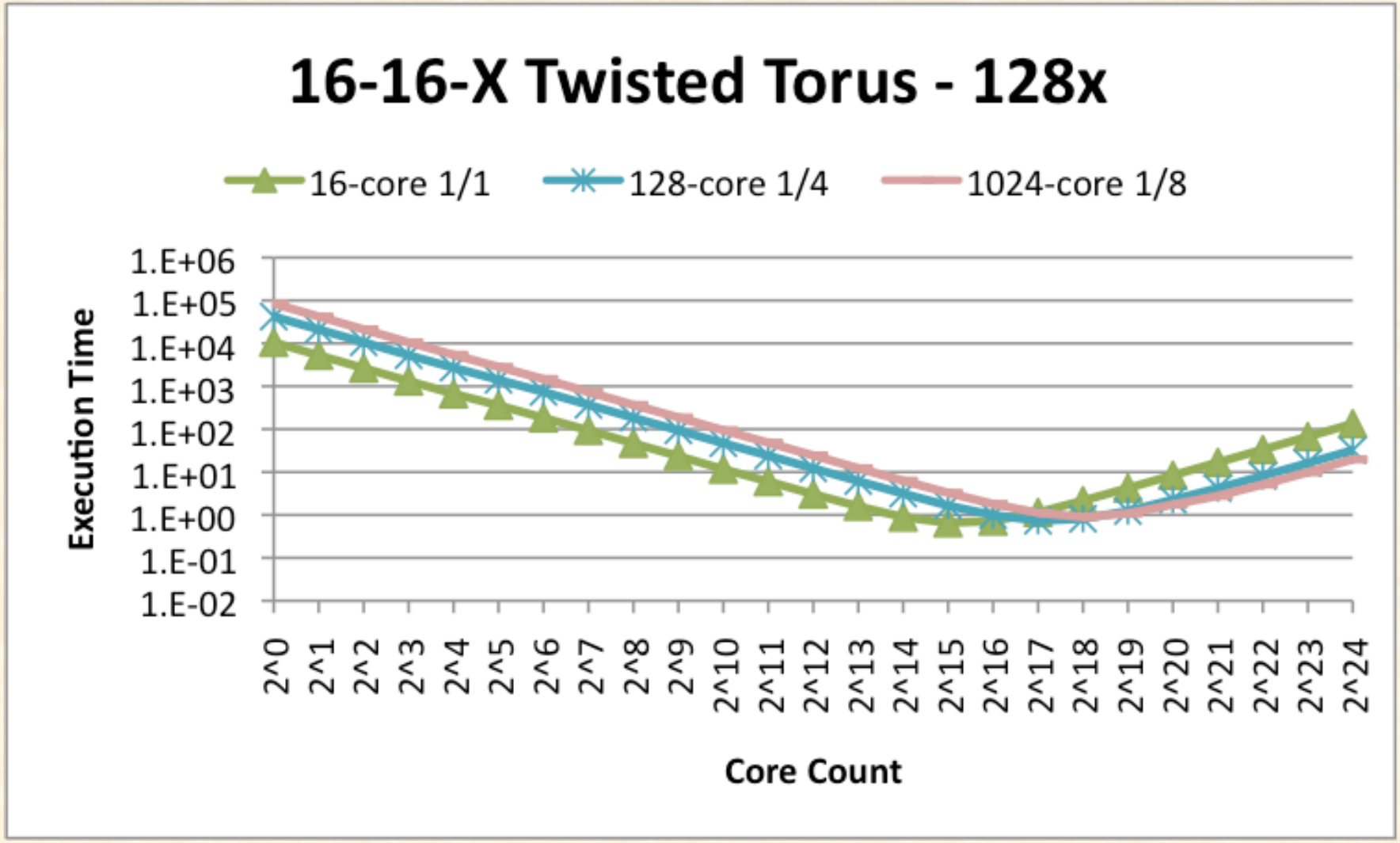
C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Scaling with X-core 3D/16 Twisted Torus (X AMD Opteron CPU, Cray XT5 Network)



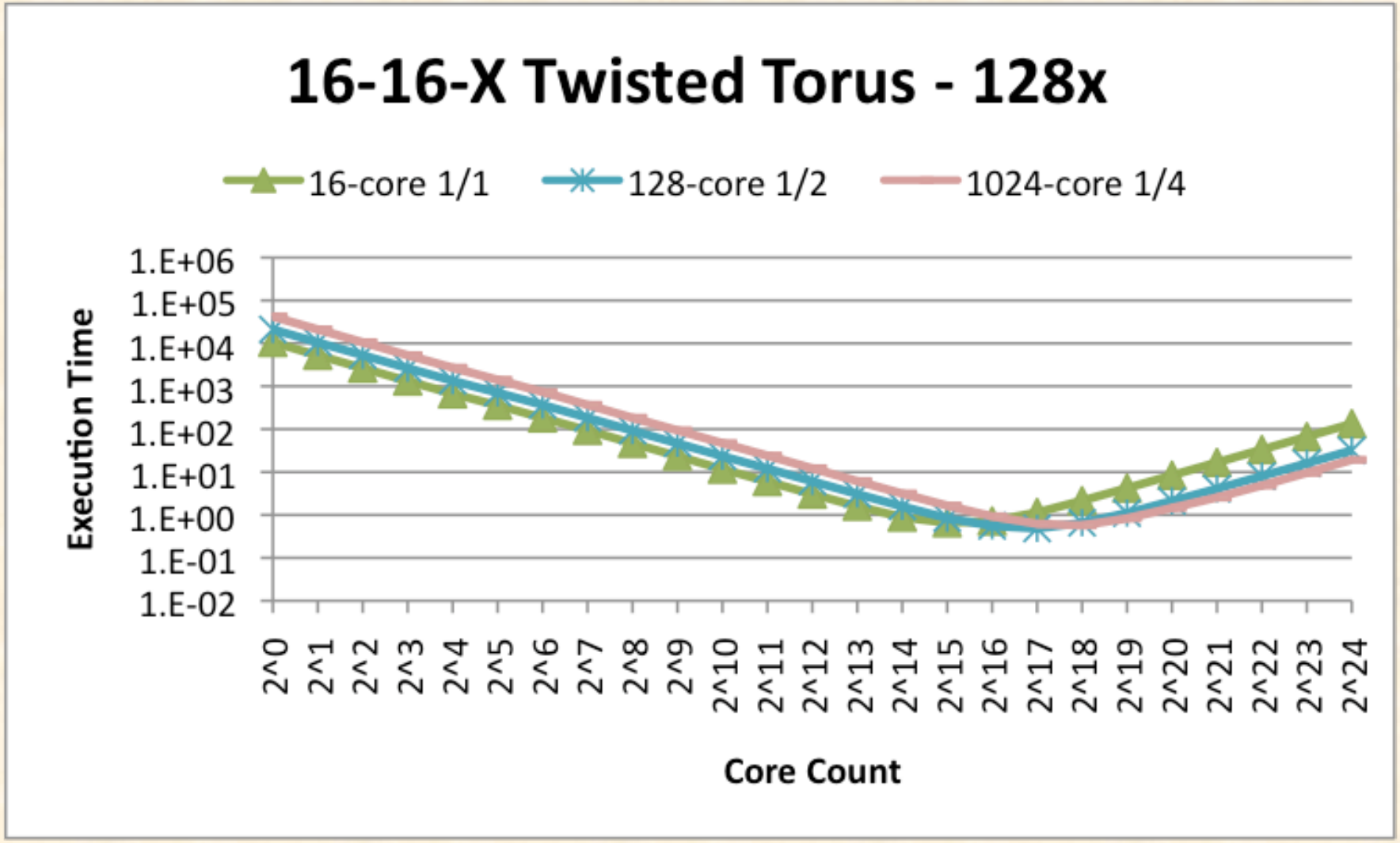
C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Scaling with X-core 3D/16 Twisted Torus (X AMD Opteron CPU, Cray XT5 Network)



C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Scaling with X-core 3D/16 Twisted Torus (X AMD Opteron CPU, Cray XT5 Network)



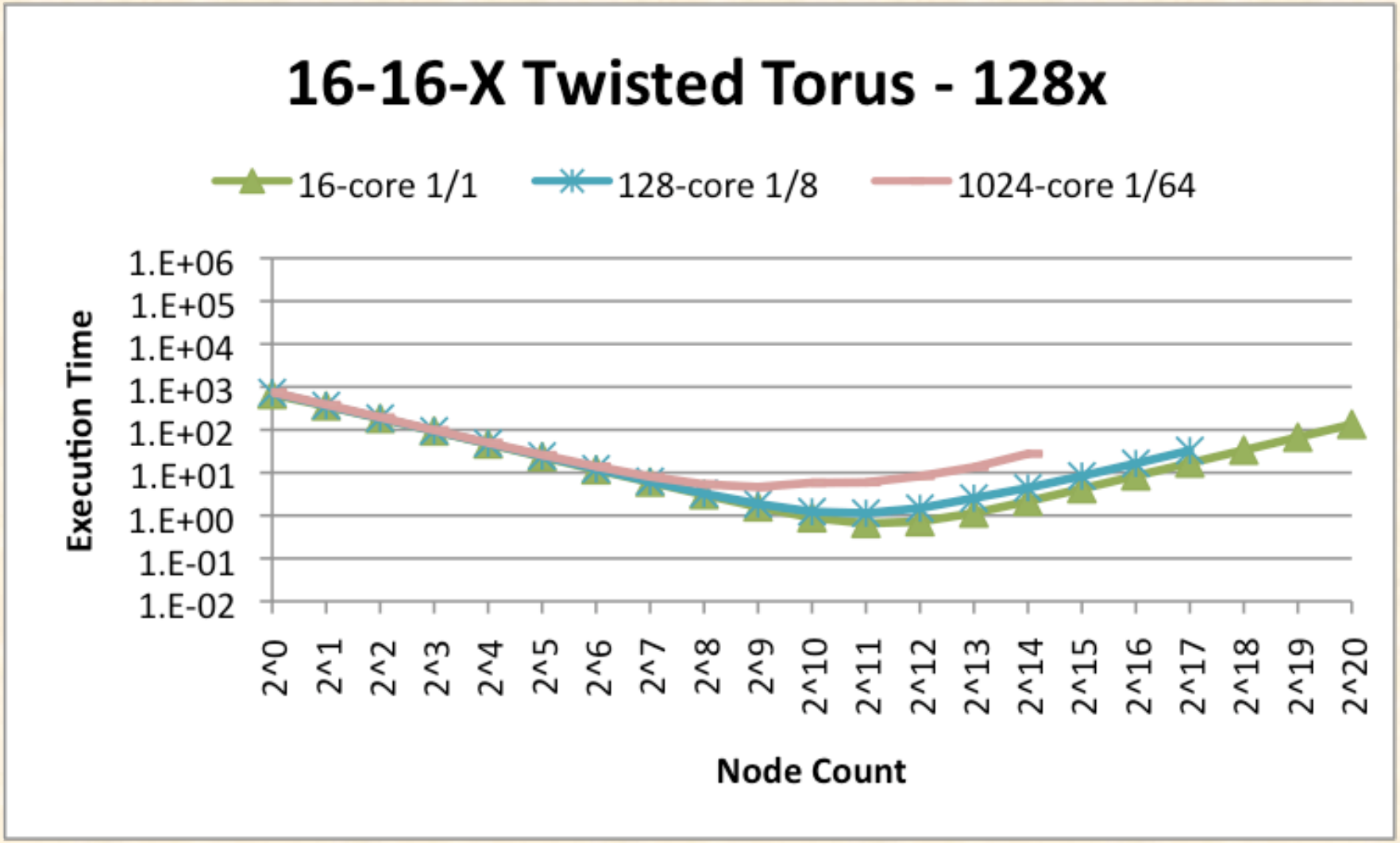
C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Scaling Properties with Less Powerful Cores

- Perfect network
  - According to Amdahl's law, the same linear scaling until 2^18 (262,144) cores independent of core performance !!!
- 3-D 16x16x.. Twisted torus with performance loss equivalent to core density (1/8 and 1/64)
  - 16-core performs best at 2^15 (32,768),
    128-core is worse with best performance at 2^18 (262,144),
    1024-core is worst with best performance at 2^19 (524,288)
- 3-D 16x16x.. twisted torus with 1/4 and 1/8 losses
  - Performance differences less, but still noticeable
  - 128- and 1024-core have best performance at 2^17 and 2^18
- 3-D 16x16x.. twisted torus with 1/2 and 1/4 losses
  - Difference is even less with bests at ~2^17 for 128/1024

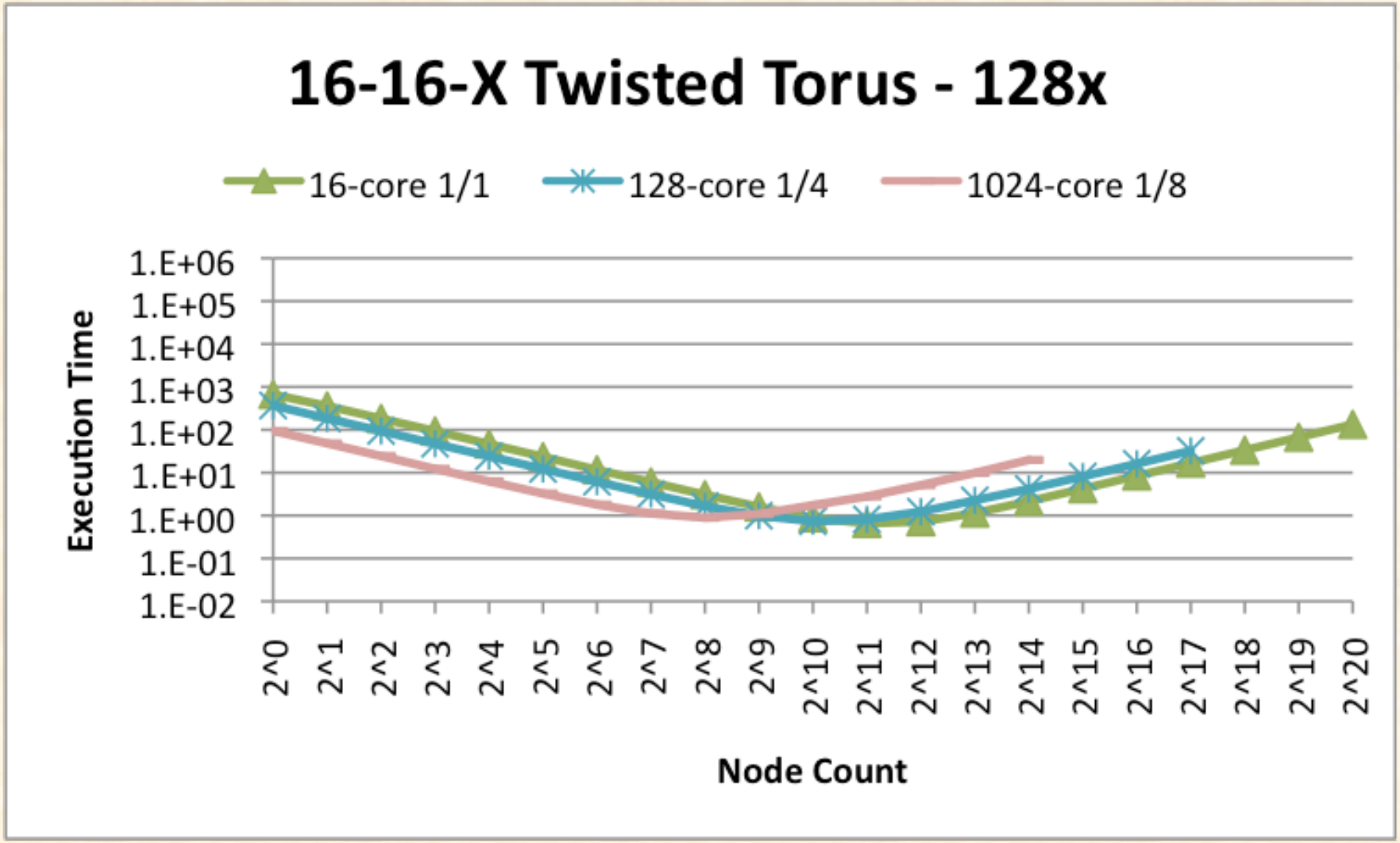
C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Scaling with X-core 3D/16 Twisted Torus (X AMD Opteron CPU, Cray XT5 Network)



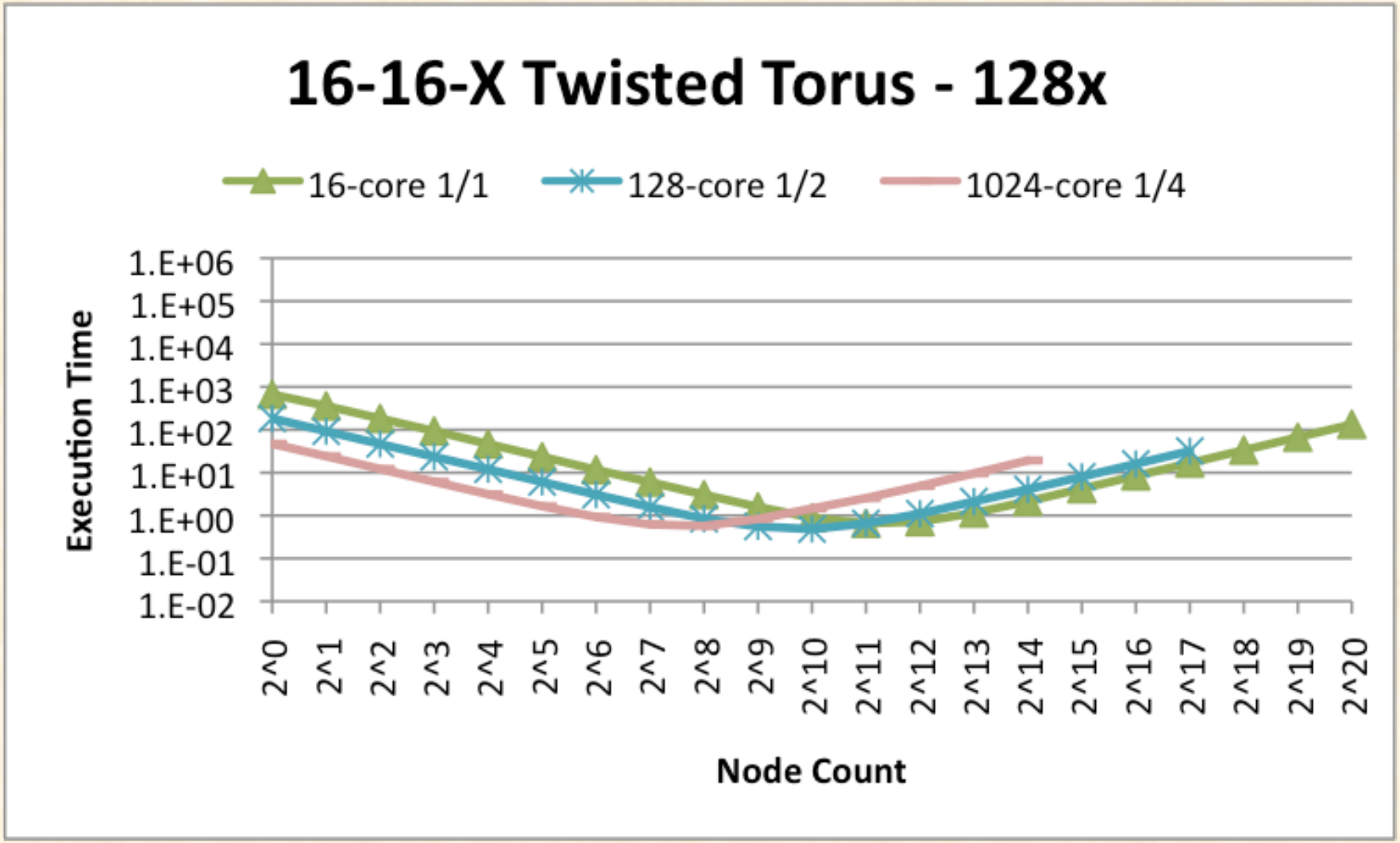
C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Scaling with X-core 3D/16 Twisted Torus (X AMD Opteron CPU, Cray XT5 Network)



C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Scaling with X-core 3D/16 Twisted Torus (X AMD Opteron CPU, Cray XT5 Network)



C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Node Scaling Properties with Less Powerful Cores

- Mirrors core scaling results
- Demonstrates node-count difference with different cores-per-processor in 3D twisted torus
- Shows equal node performance for 3-D 16x16x.. twisted torus with performance loss equivalent to core density
- Presents the clear indication that Amdahl's law will be one of the biggest challenges for exascale systems
- Shows that decreasing core performance with increasing core density is a scaling problem by itself

C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

Conclusions and the Path Forward

- Strong scaling algorithms to millions of cores will be a clear challenge due to Amdahl's law
- The expected performance loss with increasing core density is exacerbating this problem
- Applications typically consist of multiple sequentially executed parallel algorithms that scale differently
  - Impact of individual algorithm scalability on execution time
- Possible solutions include (moving away from SPMD)
  - Schedule different application phases in parallel (considering data dependency, on same or different nodes)
  - Temporarily repartition application data to smaller scale (considering costs for data movement, shrink/grow app.)
  - Replace algorithms or create hybrid algorithms

Future Work

- Investigate different algorithms in a comparative study to show different application and system scaling properties
- Improve simulation capabilities with basic power model to include the design envelope for power consuption
- Enhance simulator with instrumentation for processor/ memory usage modeling using performance counters
- Create application models using instrumentation and execute application models instead of applications
- Model network congestion in a scalable fashion
- Model file I/O systems, e.g, to discover I/O congestion
- Resilience/performance co-design using fault injection

C. Engelmann. Scaling To A Million Cores And Beyond: A Basic Understanding Of The Challenges Ahead On The Road To Exascale.

UT-BATTELLE

Questions?