GPU COMPUTING TO EXASCALE AND BEYOND

BILL DALLY
CHIEF SCIENTIST & SVP OF RESEARCH, NVIDIA

GPU Computing

1. The GPU Advantage
2. To ExaScale and Beyond
3. The GPU is the Computer

The GPU Advantage

The GPU Advantage

A Tale of Two Machines

Tianhe-1A

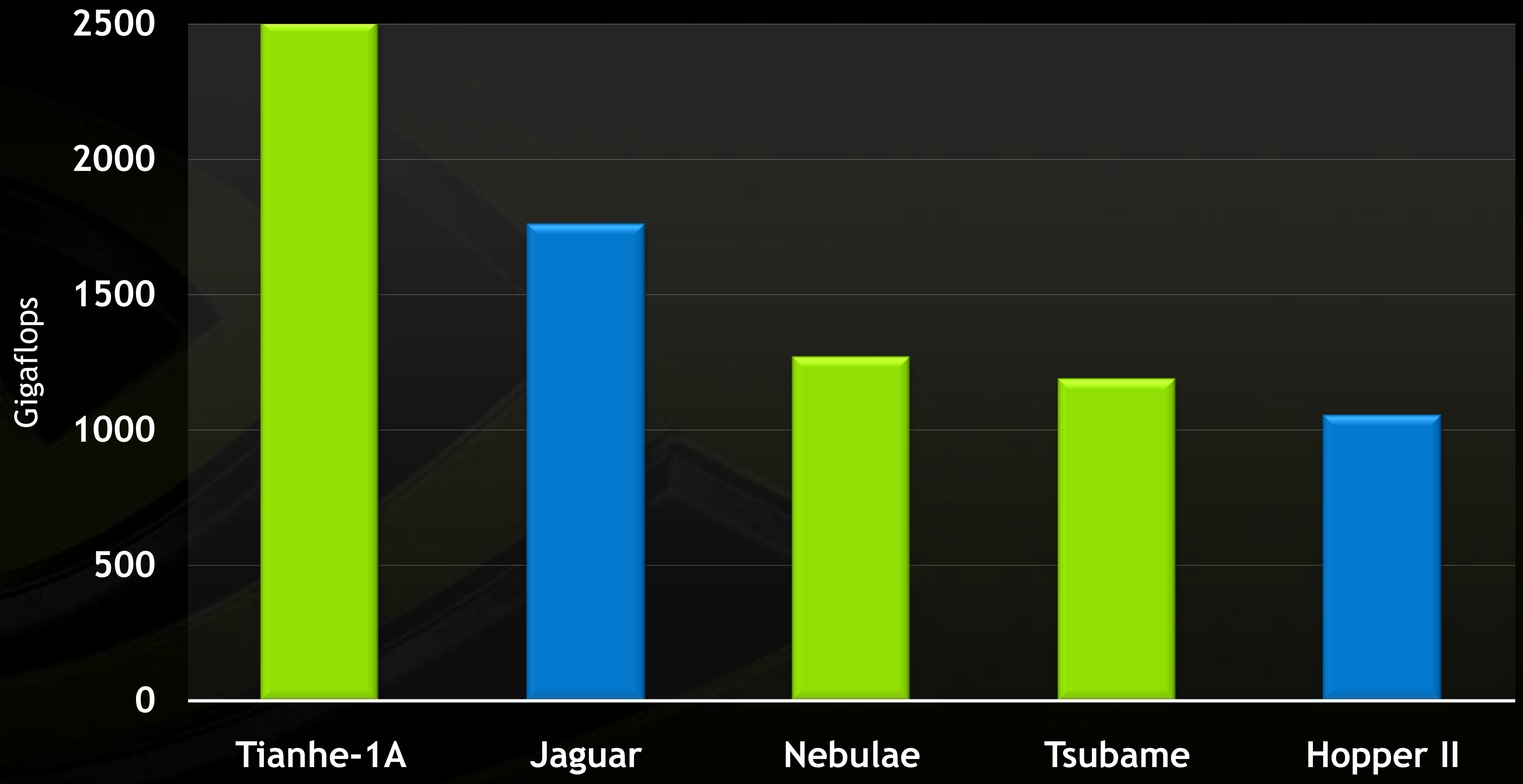
at NSC Tianjin

Tianhe-1A
at NSC Tianjin

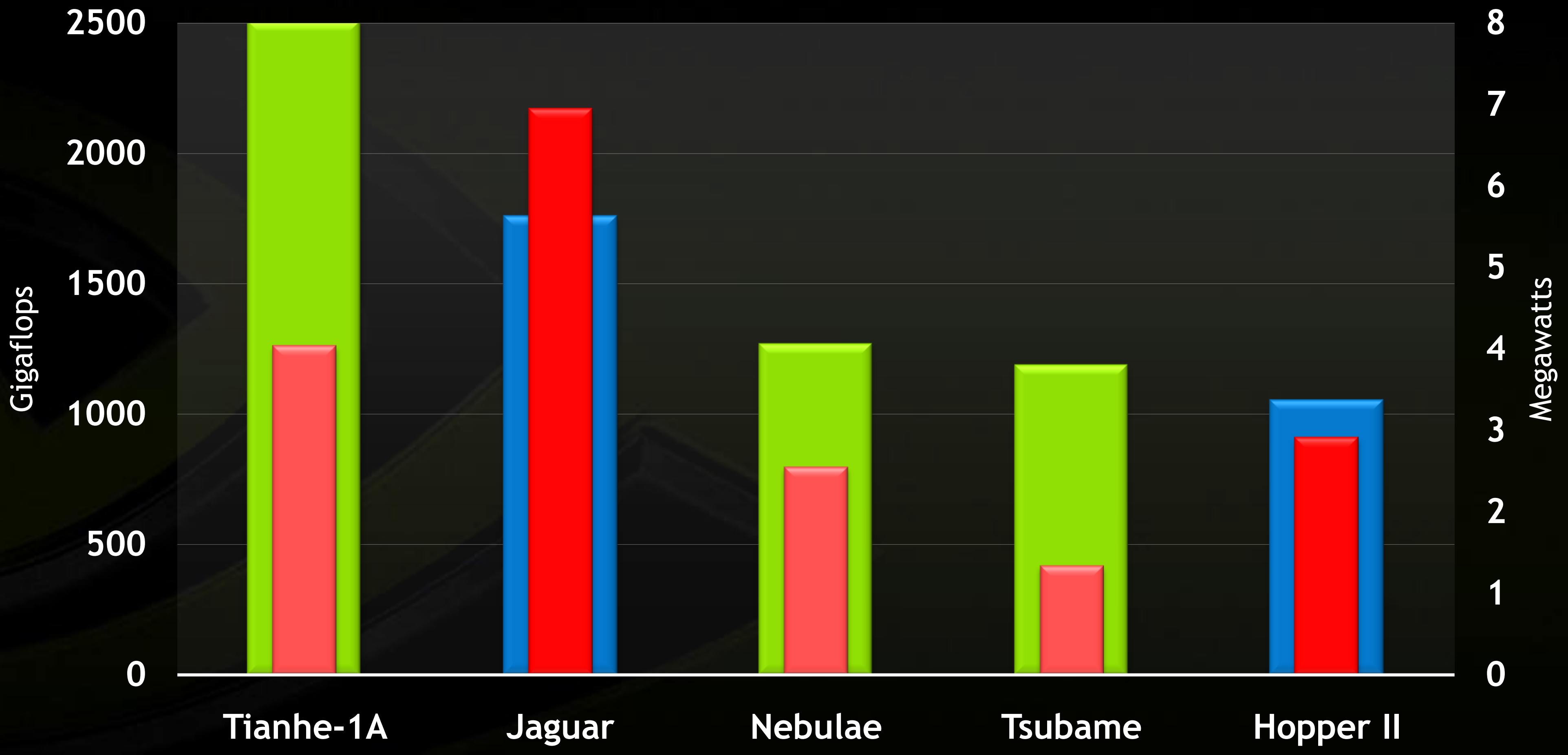
- The World's Fastest Supercomputer
- 2.507 Petaflop
- 7168 Tesla M2050 GPUs

Tesla M2050 GPUs

3 of Top5 Supercomputers



Top 5 Performance and Power

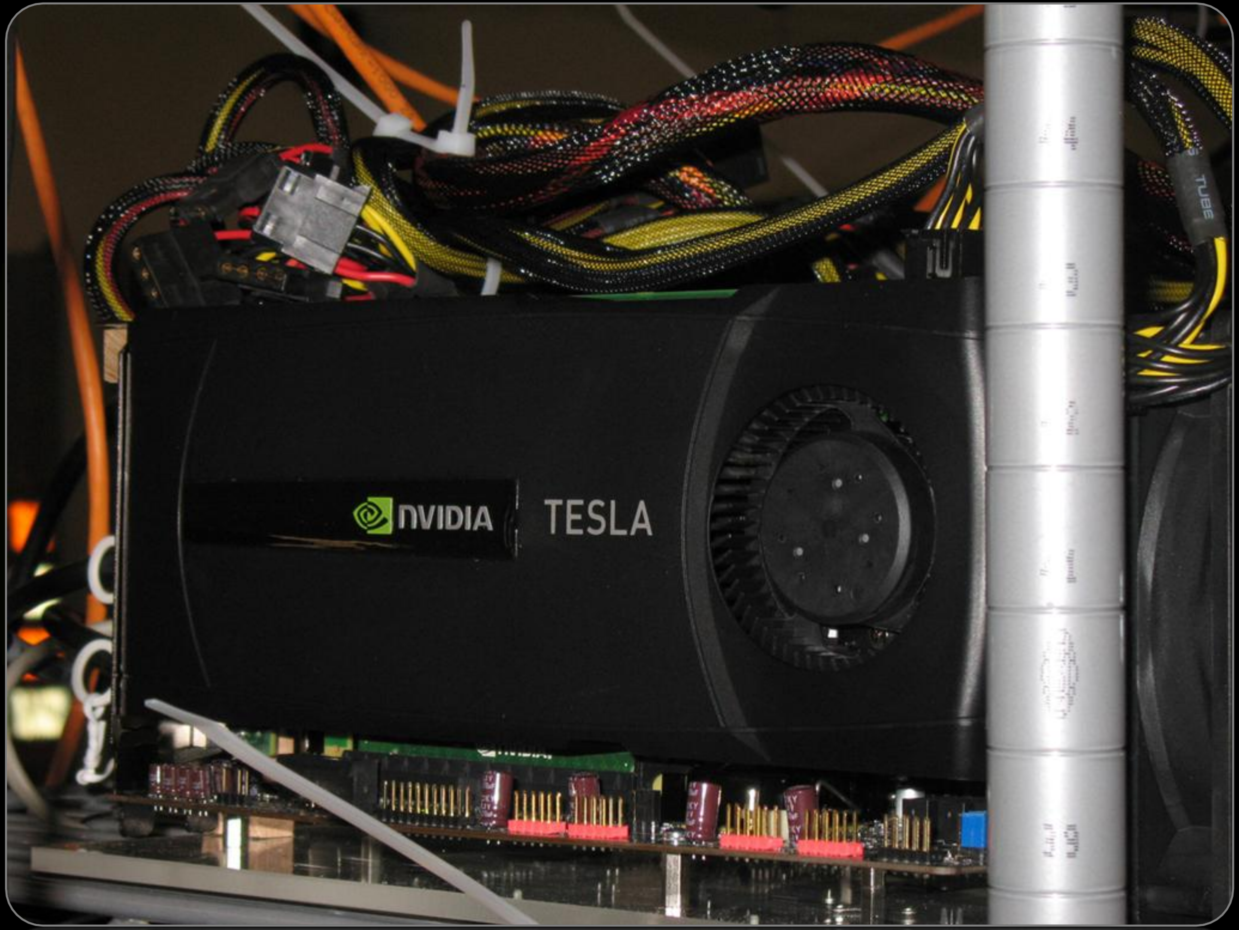


NVIDIA/NCSA

Green 500 Entry

NVIDIA/NCSA

Green 500 Entry



NVIDIA/NCSA Green 500 Entry

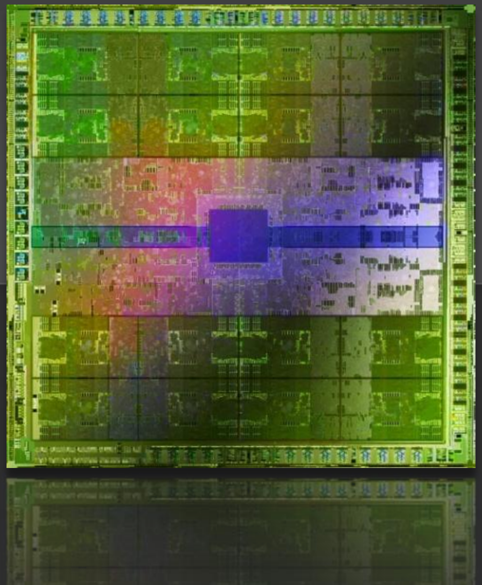
- **128 nodes, each with:**
  - 1x Core i3 530 (2 cores, 2.93 GHz => 23.4 GFLOP peak)
  - 1x Tesla C2050 (14 cores, 1.15 GHz => 515.2 GFLOP peak)
  - 4x QDR Infiniband
  - 4 GB DRAM
- **Theoretical Peak Perf: 68.95 TF**
- **Footprint: ~20 ft^2 => 3.45 TF/ft^2**
- **Cost: $500K (street price) => 137.9 MF/$**
- **Linpack: 33.62 TF, 36.0 kW => 934 MF/W**

The GPU Advantage

Efficiency and Programmability

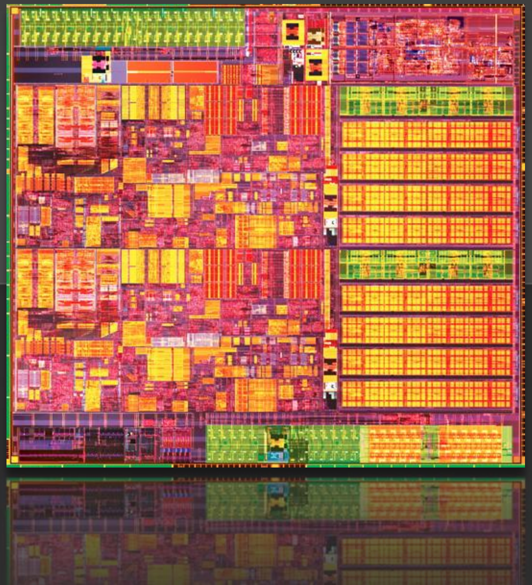
GPU

200pJ/Instruction



CPU

2nJ/Instruction

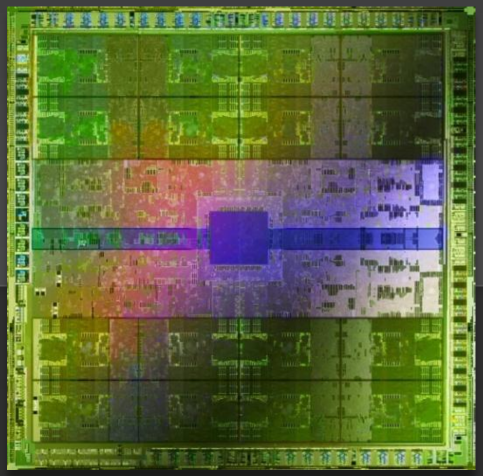


GPU

200pJ/Instruction

Optimized for Throughput

Explicit Management
of On-chip Memory

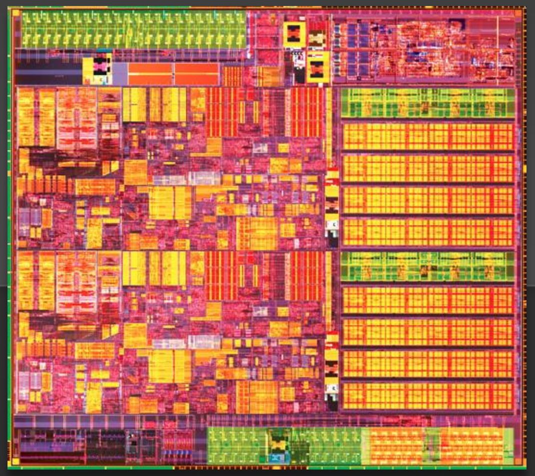


CPU

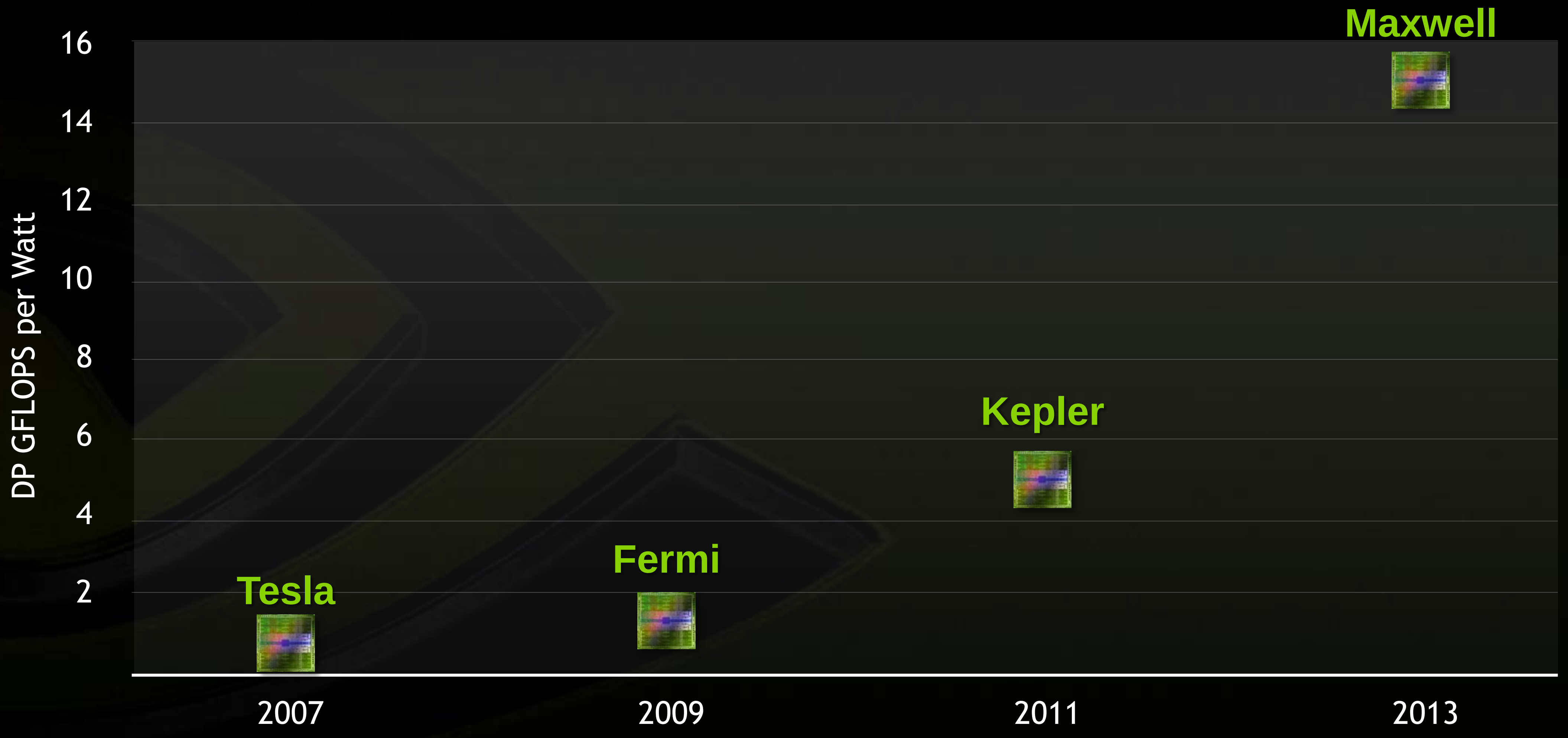
2nJ/Instruction

Optimized for Latency

Caches



CUDA GPU Roadmap



The GPU Advantage

Efficiency and Programmability

The GPU Advantage

CUDA Enables Programmability

CUDA C: C with a Few Keywords

```
void saxpy_serial(int n, float a, float *x, float *y)
{
    for (int i = 0; i < n; ++i)
        y[i] = a*x[i] + y[i];
}
// Invoke serial SAXPY kernel
saxpy_serial(n, 2.0, x, y);
```

Standard C Code

```
__global__ void saxpy_parallel(int n, float a, float *x, float *y)
{
    int i = blockIdx.x*blockDim.x + threadIdx.x;
    if (i < n)  y[i] = a*x[i] + y[i];
}
// Invoke parallel SAXPY kernel with 256 threads/block
int nblocks = (n + 255) / 256;
saxpy_parallel<<<nblocks, 256>>>(n, 2.0, x, y);
```

CUDA C Code

GPU Computing Ecosystem

DESIGNED FOR NVIDIA® CUDA™

Libraries

$$\oint \mathbf{E} \cdot d\mathbf{A} = \frac{q_{enc}}{\varepsilon_0}$$
$$\oint \mathbf{B} \cdot d\mathbf{A} = 0$$
$$\oint \mathbf{E} \cdot d\mathbf{s} = -\frac{d\Phi_B}{dt}$$
$$\oint \mathbf{B} \cdot d\mathbf{s} = \mu_0 \varepsilon_0 \frac{d\Phi_E}{dt} + \mu_0 i_{enc}$$

Mathematical Packages

Wolfram Mathematica

LabVIEW

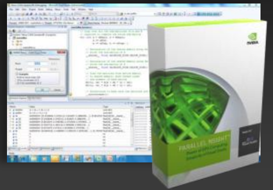
MATLAB 7

Research & Education

Consultants, Training & Certification

Integrated Development Environment

Parallel Nsight for MS Visual Studio



Microsoft Visual Studio PARTNER

Tools & Partners

PGI CDK Cluster Development Kit Software

CAPS hmpp

TOTALVIEW TECHNOLOGIES A Rogue Wave Software Company

allinea ddt the distributed debugging tool

Languages & API's

CUDA C/C++

Microsoft DirectX 11

PGI WORKSTATION Fortran

OpenCL

All Major Platforms

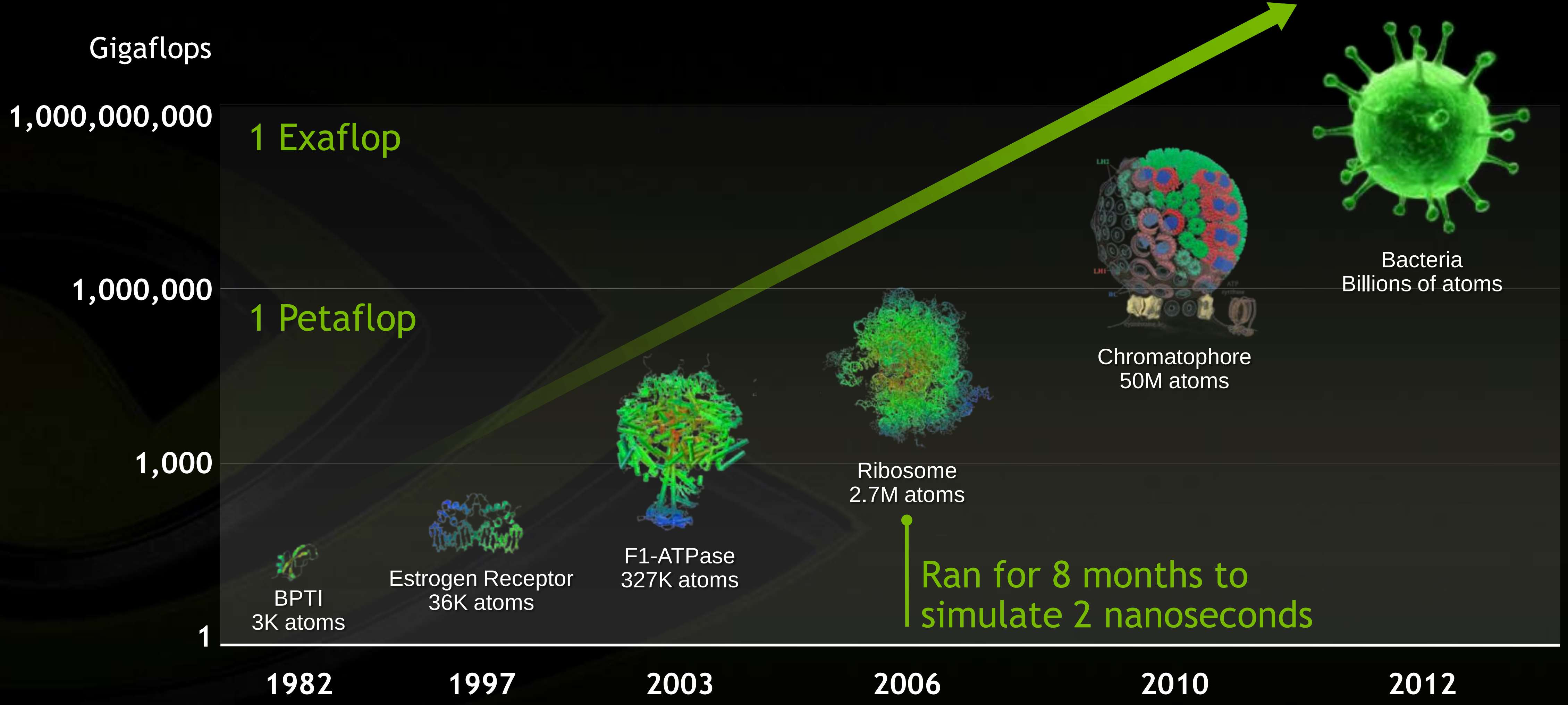
GPU Computing Today

By the Numbers:

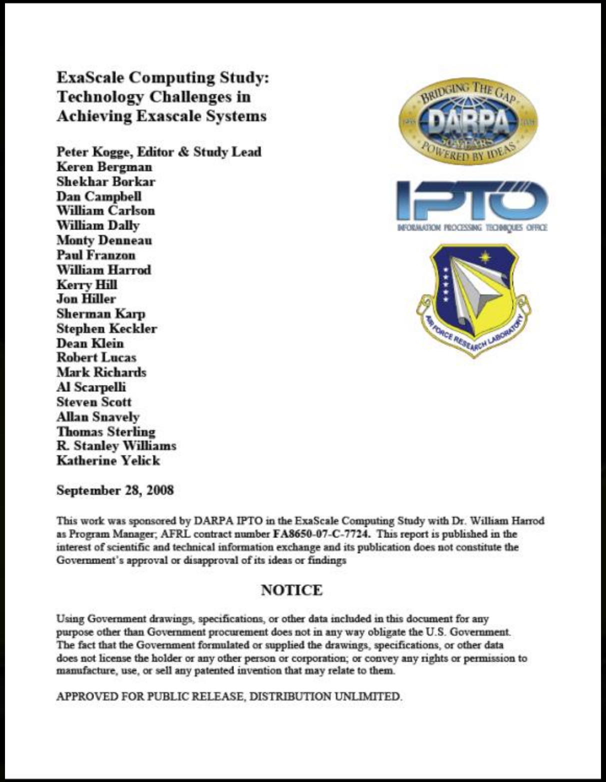
200 Million	CUDA Capable GPUs
600,000	CUDA Toolkit Downloads
100,000	Active GPU Computing Developers
8,000	Members in Parallel Nsight Developer Program
362	Universities Teaching CUDA Worldwide
11	CUDA Centers of Excellence Worldwide

To ExaScale and Beyond

Science Needs 1000x More Computing



DARPA Study Identifies Four Challenges for ExaScale Computing



ExaScale Computing Study:
Technology Challenges in
Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number FA8650-07-C-7724. This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings

NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.

Available at
www.darpa.mil/ipto/personnel/docs/ExaScale_Study_Initial.pdf

Report published September 28, 2008:

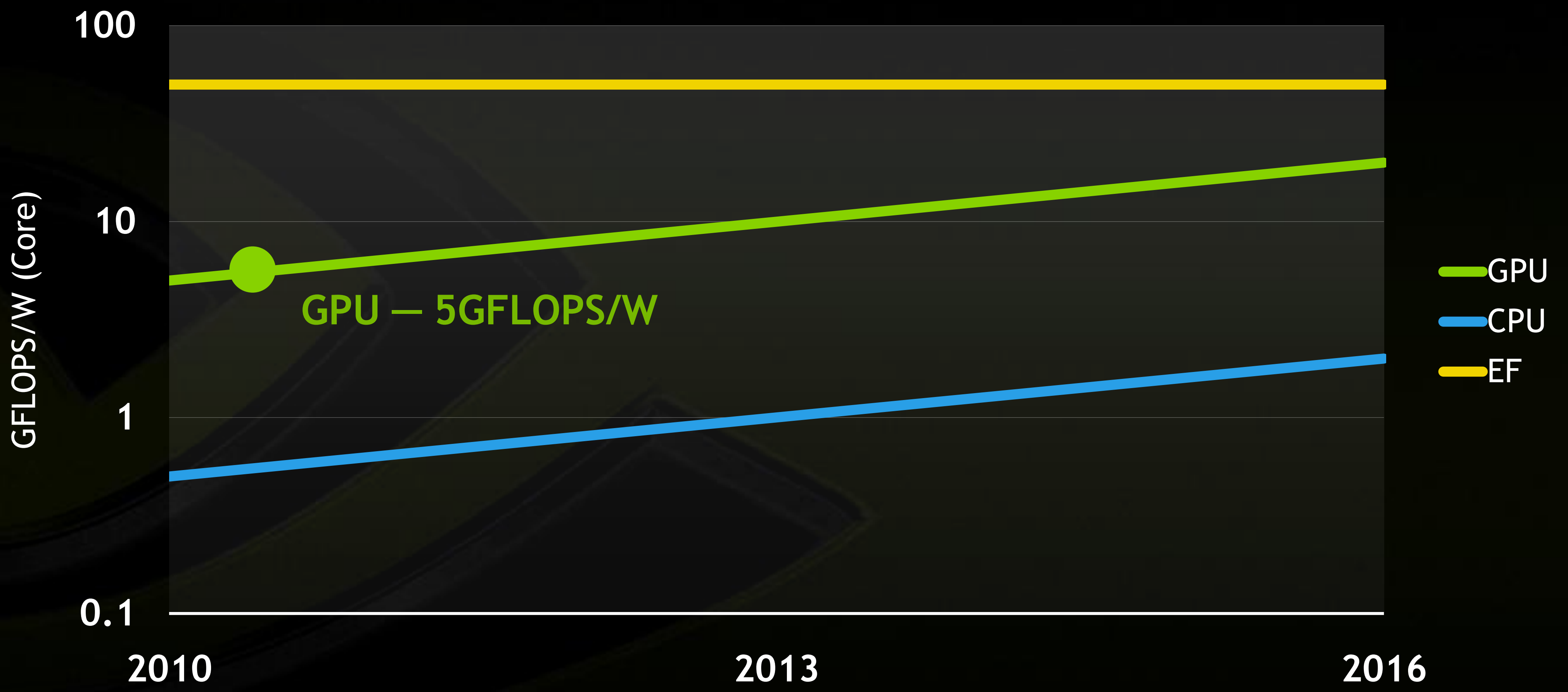
- **Four Major Challenges**
  - Energy and Power challenge
  - Memory and Storage challenge
  - Concurrency and Locality challenge
  - Resiliency challenge
- **Number one issue is *power***
  - Extrapolations of current architectures and technology indicate over 100MW for an Exaflop!
  - Power also constrains what we can put on a chip

Power is THE Problem

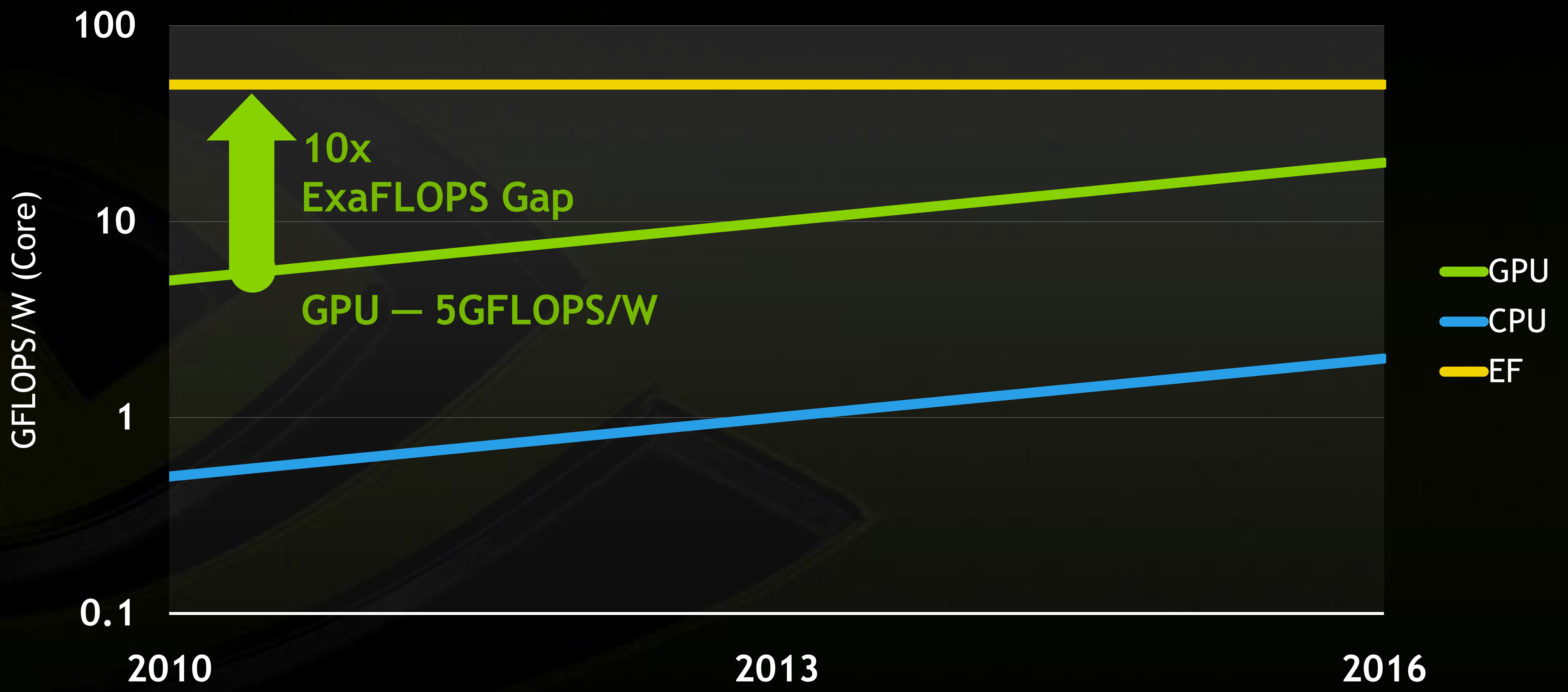
Power is THE Problem

A GPU is the Solution

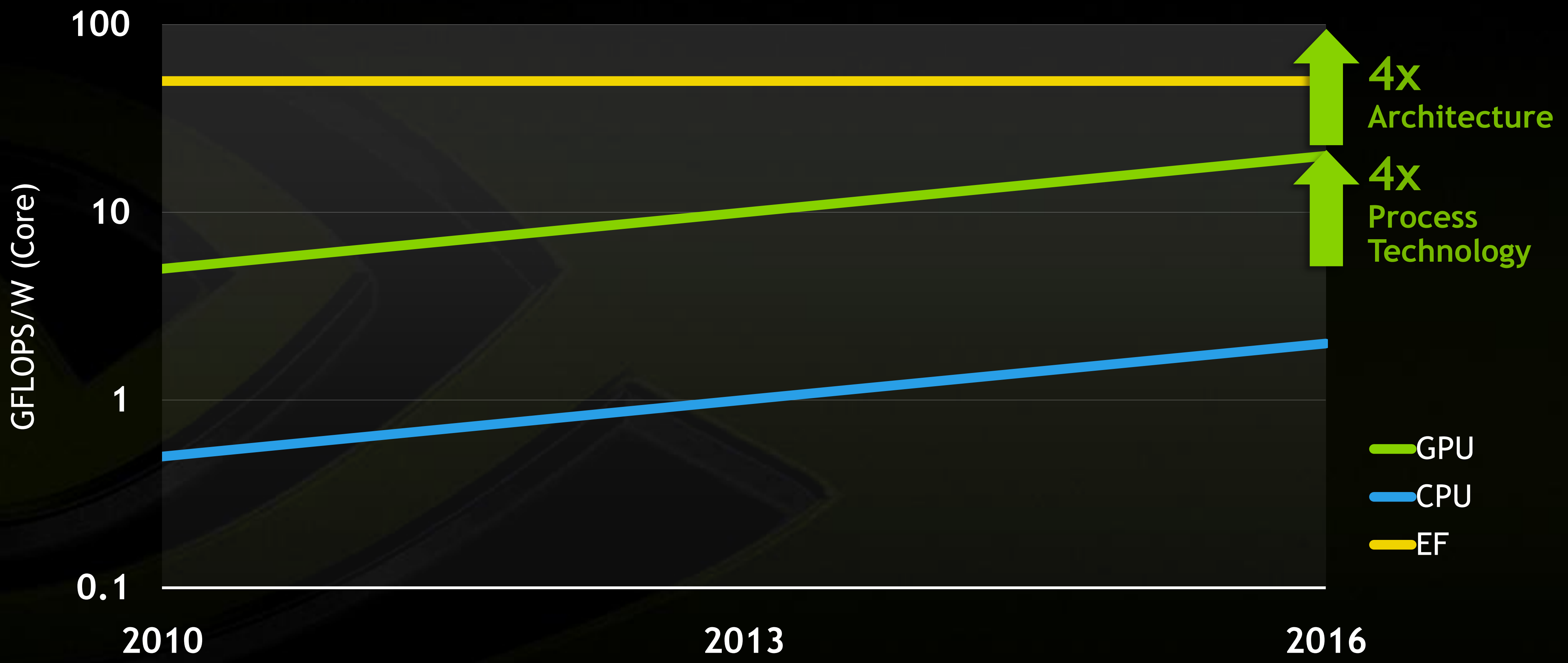
ExaFLOPS at 20MW = 50GFLOPS/W



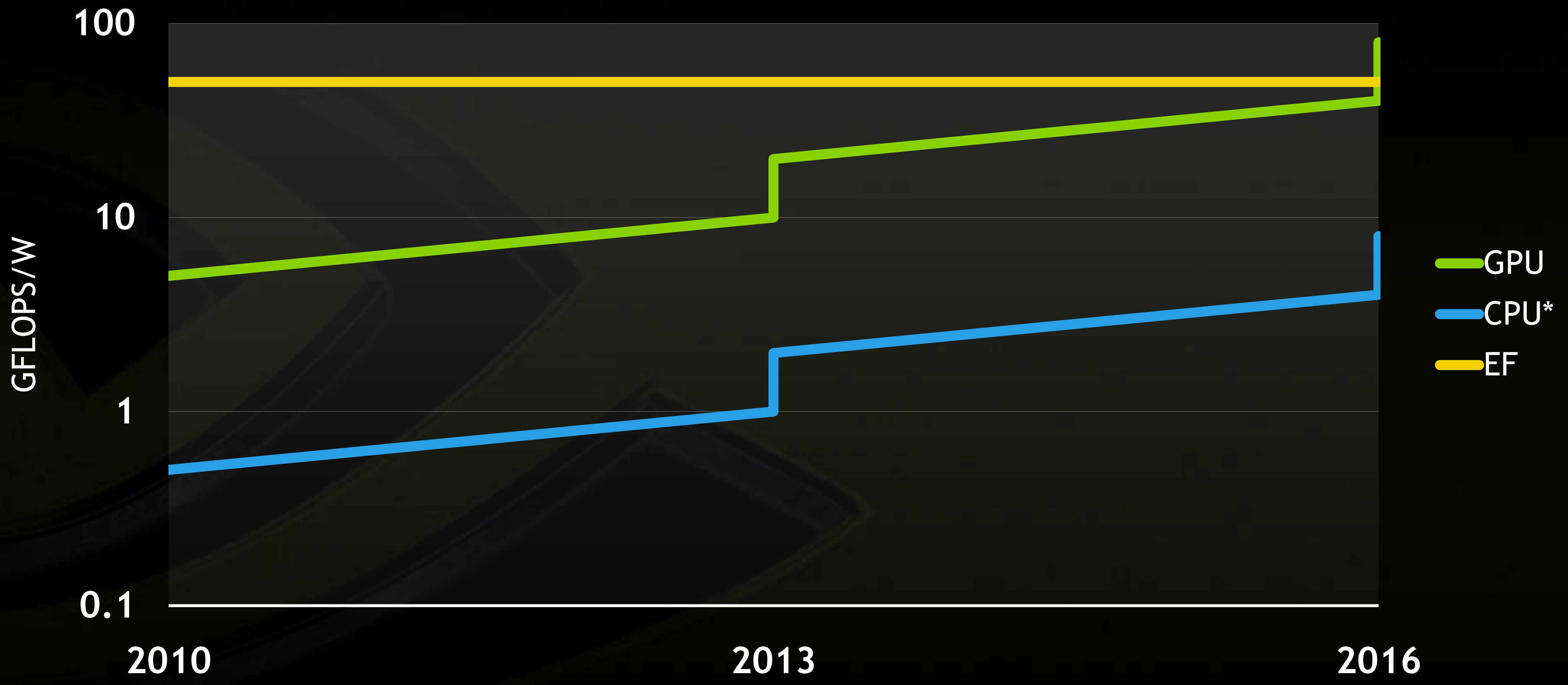
50GFLOPS/W
10x Energy Gap for Today's GPU



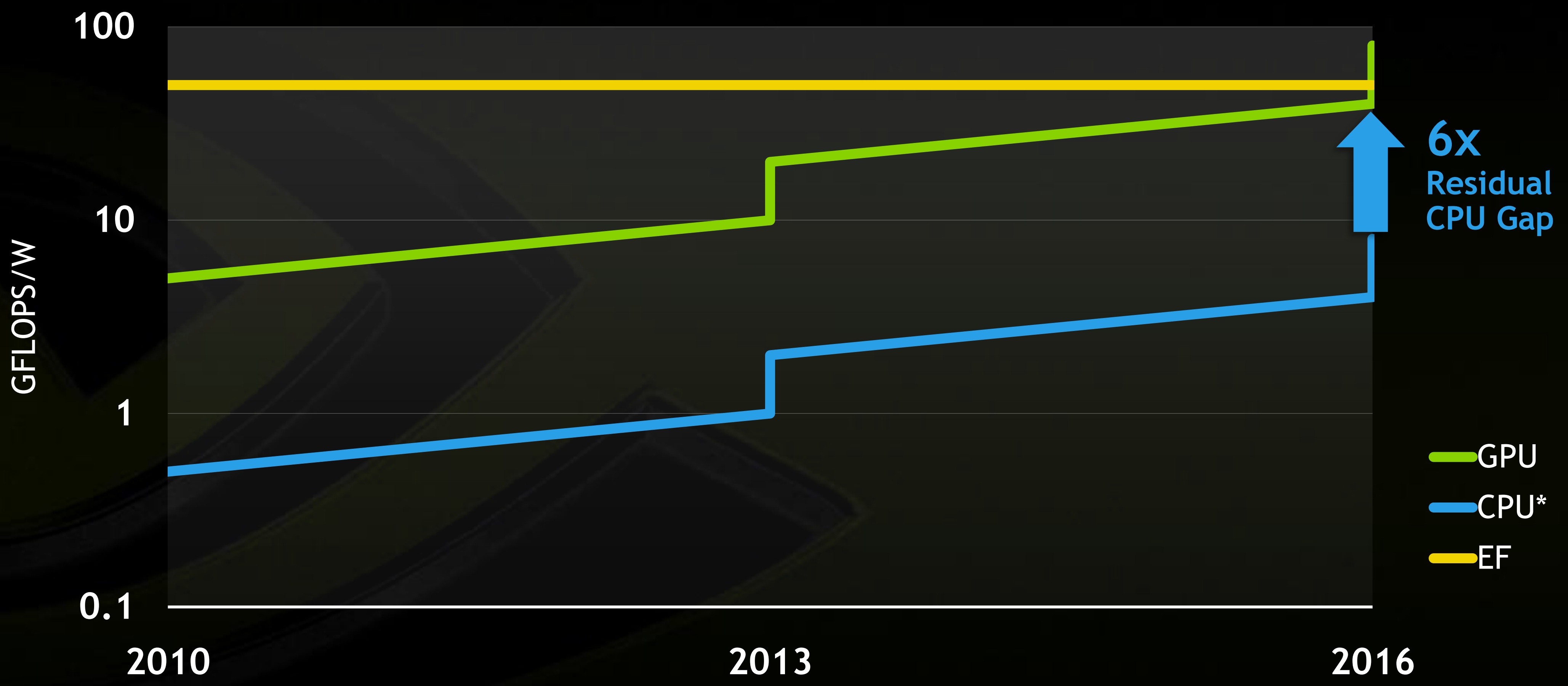
GPUs Close the Gap with Process and Architecture

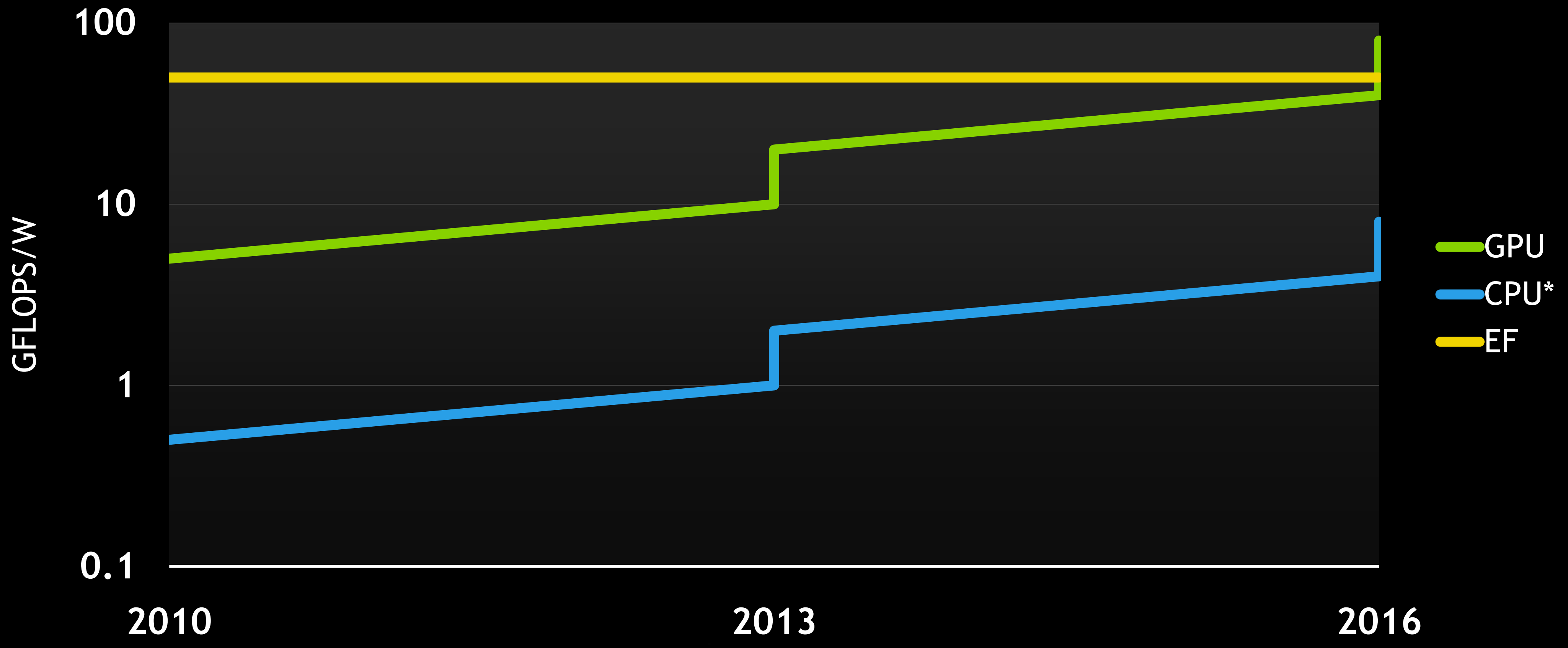


GPUs Close the Gap with Process and Architecture



GPUs Close the Gap
With CPUs, a Gap Remains





GPUs Close the Gap
With CPUs, a Gap Remains

Heterogeneous Computing
is Required to get to ExaScale

Echelon

NVIDIA’s Extreme-Scale Computing Project

Echelon Team

THE UNIVERSITY OF UTAH

NVIDIA

Penn

CRAY

OAK RIDGE National Laboratory

Micron

THE UNIVERSITY OF TEXAS AT AUSTIN

THE UNIVERSITY of TENNESSEE UT

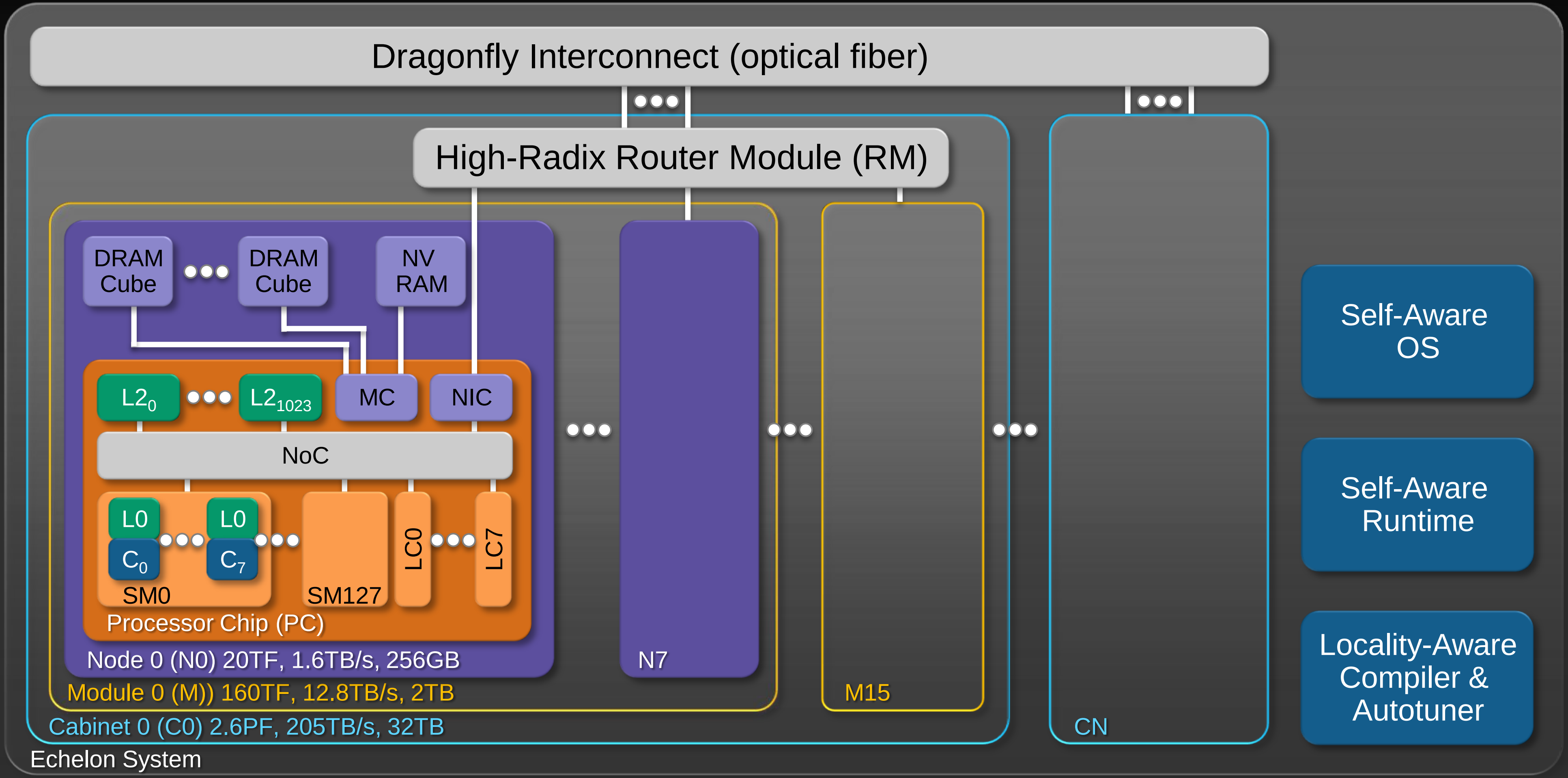
Georgia Institute of Technology

THE UNIVERSITY OF CALIFORNIA · LET THERE BE LIGHT · 1868

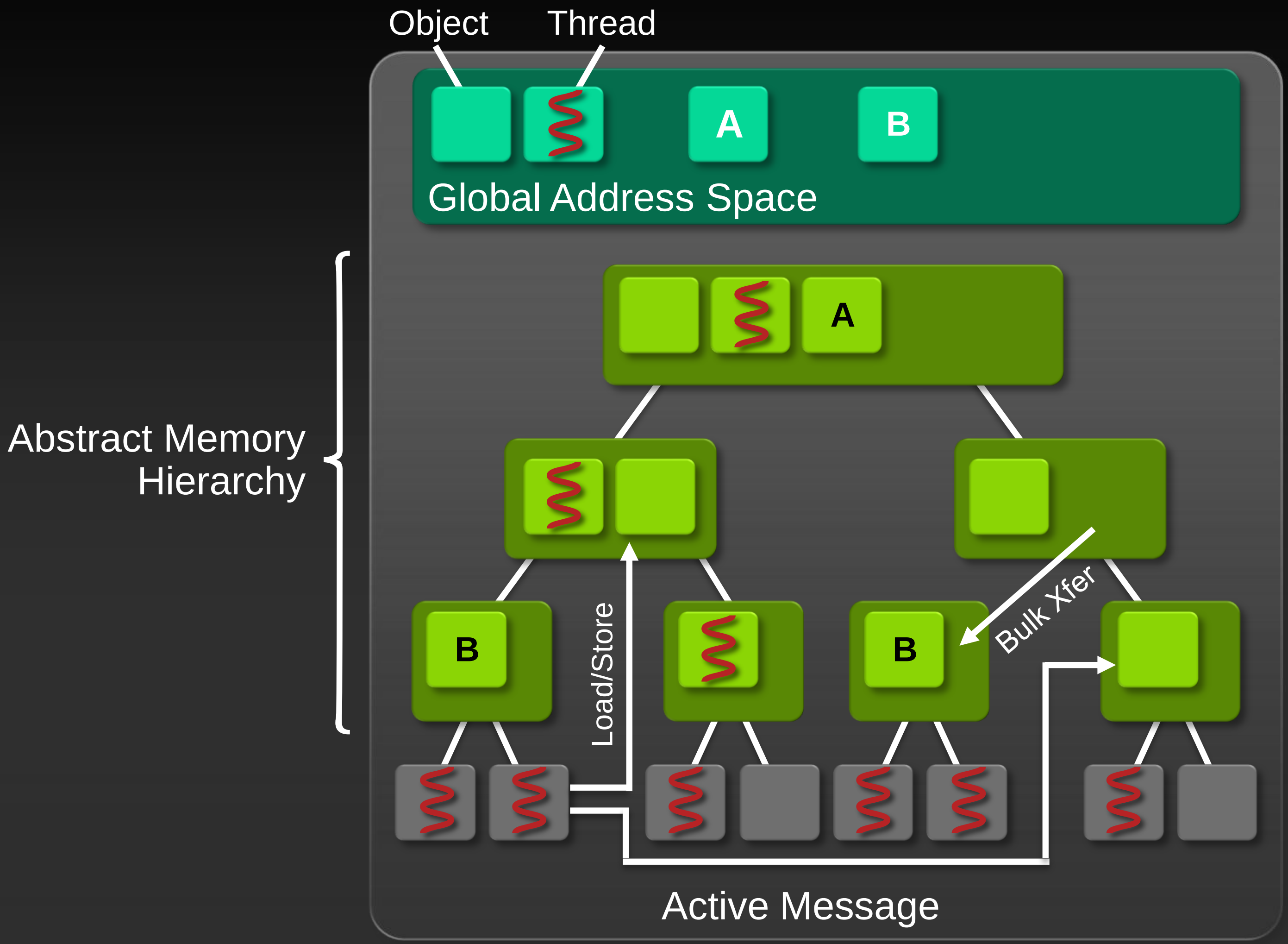
LOCKHEED MARTIN

LELAND STANFORD JUNIOR UNIVERSITY · DIE LUFT DER FREIHEIT WEHT · 1891

System Sketch

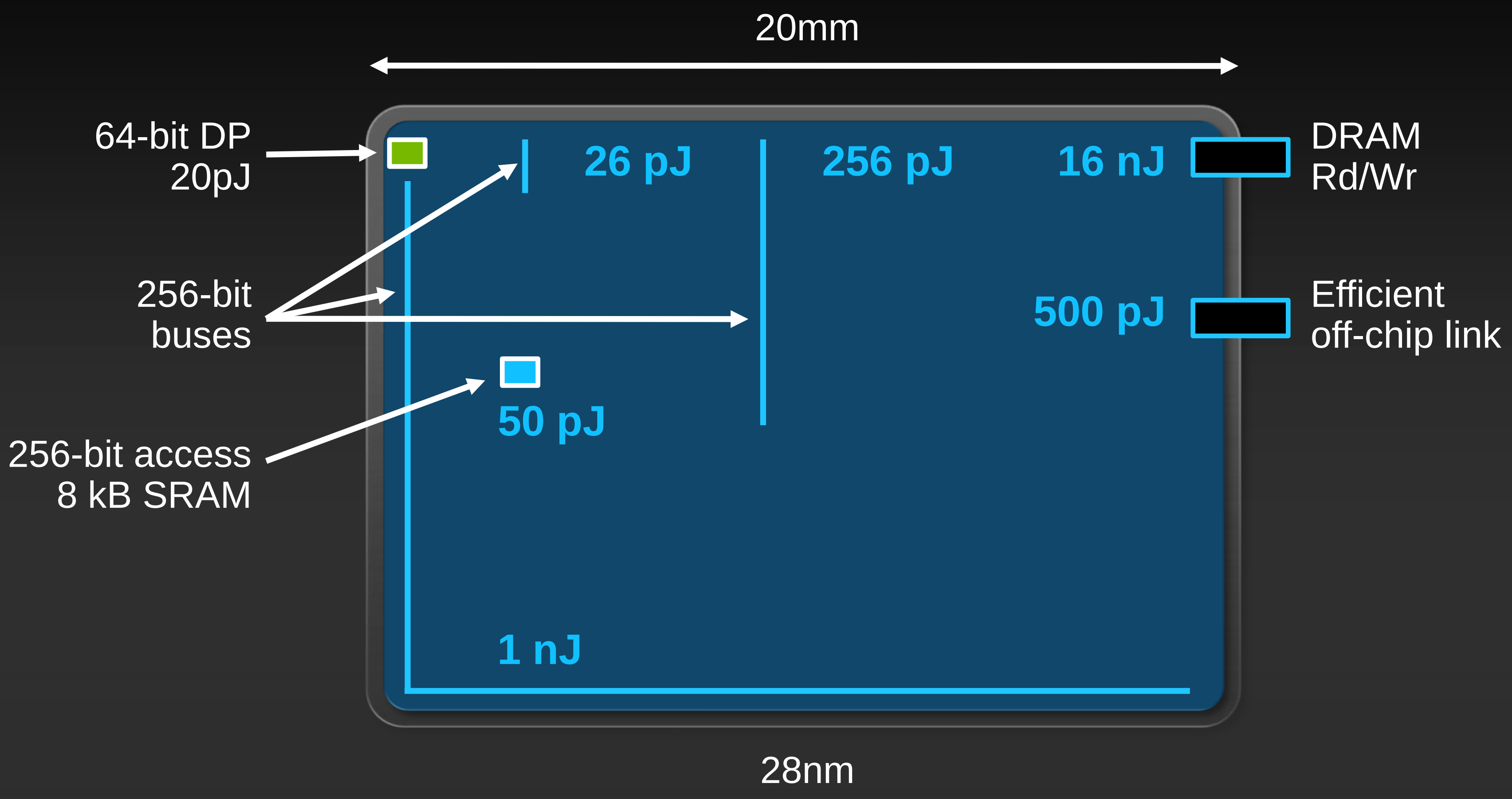


Execution Model



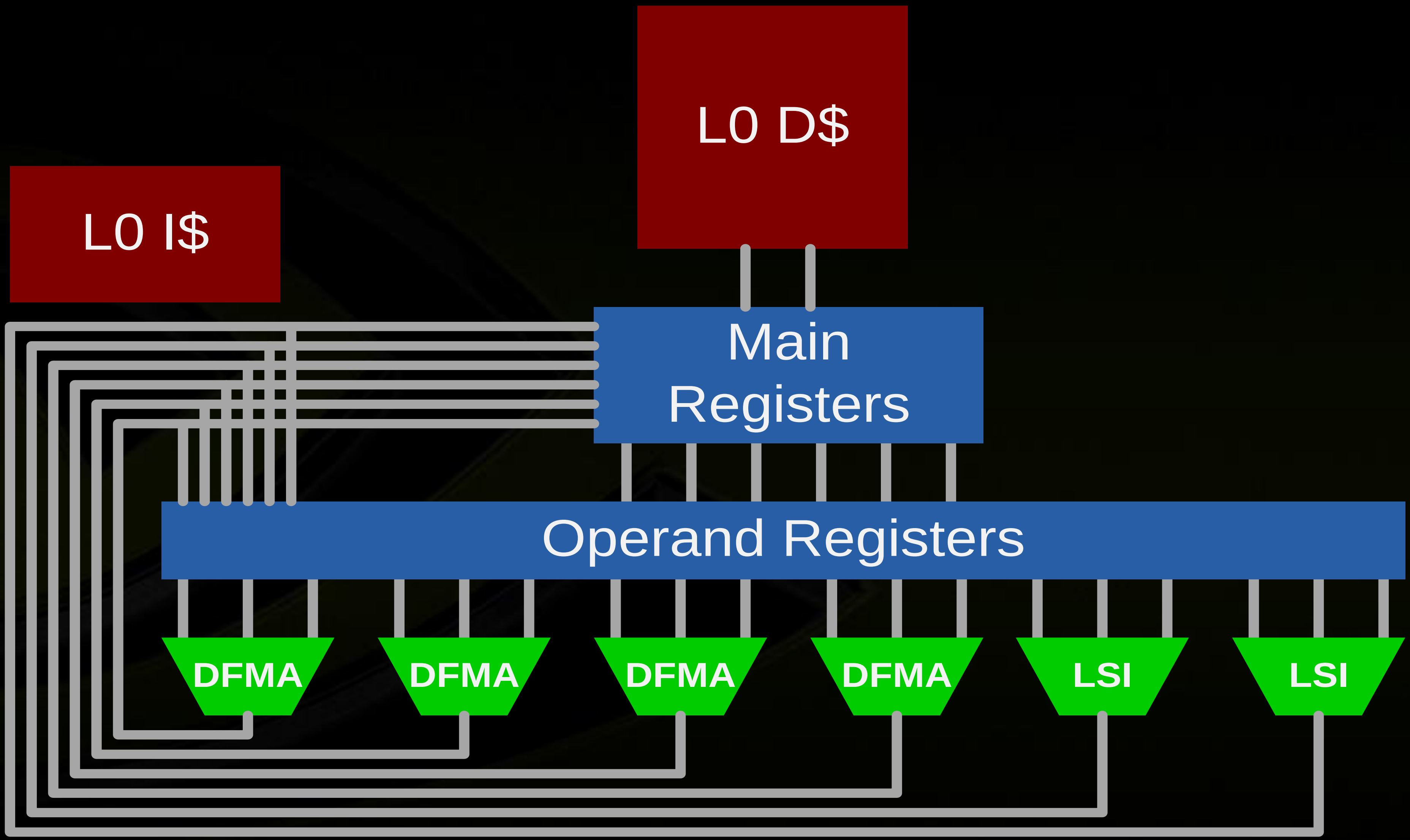
The High Cost of Data Movement

Fetching operands costs more than computing on them

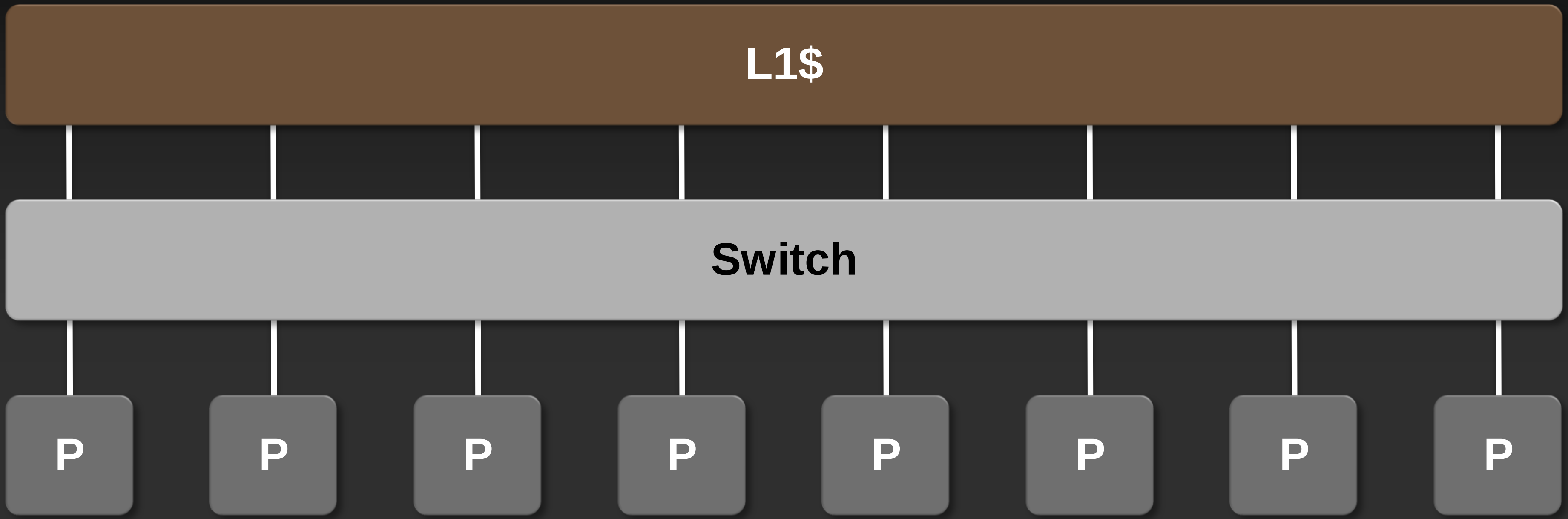


An NVIDIA ExaScale Machine

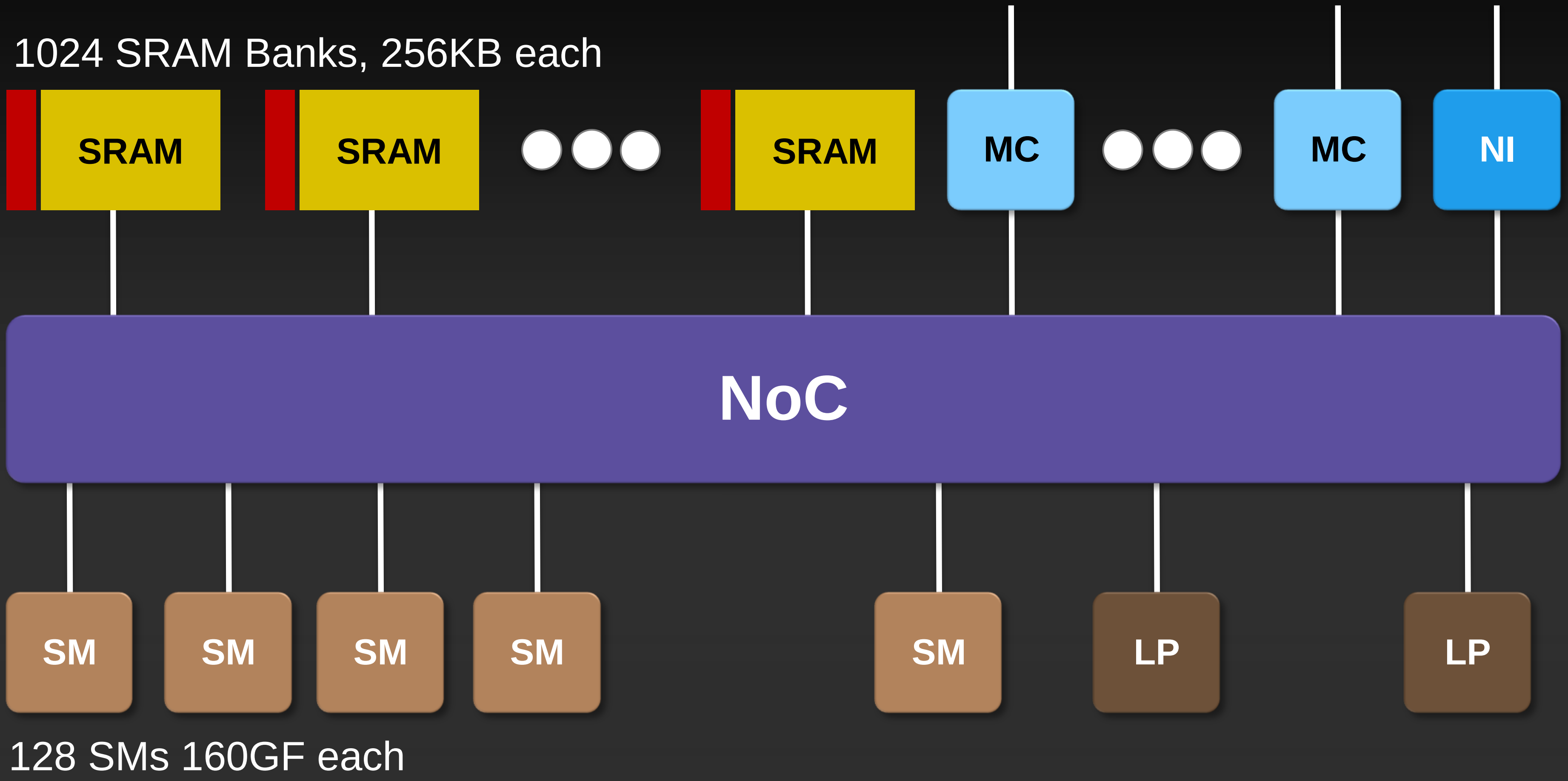
Lane – 4 DFMAs, 20GFLOPS



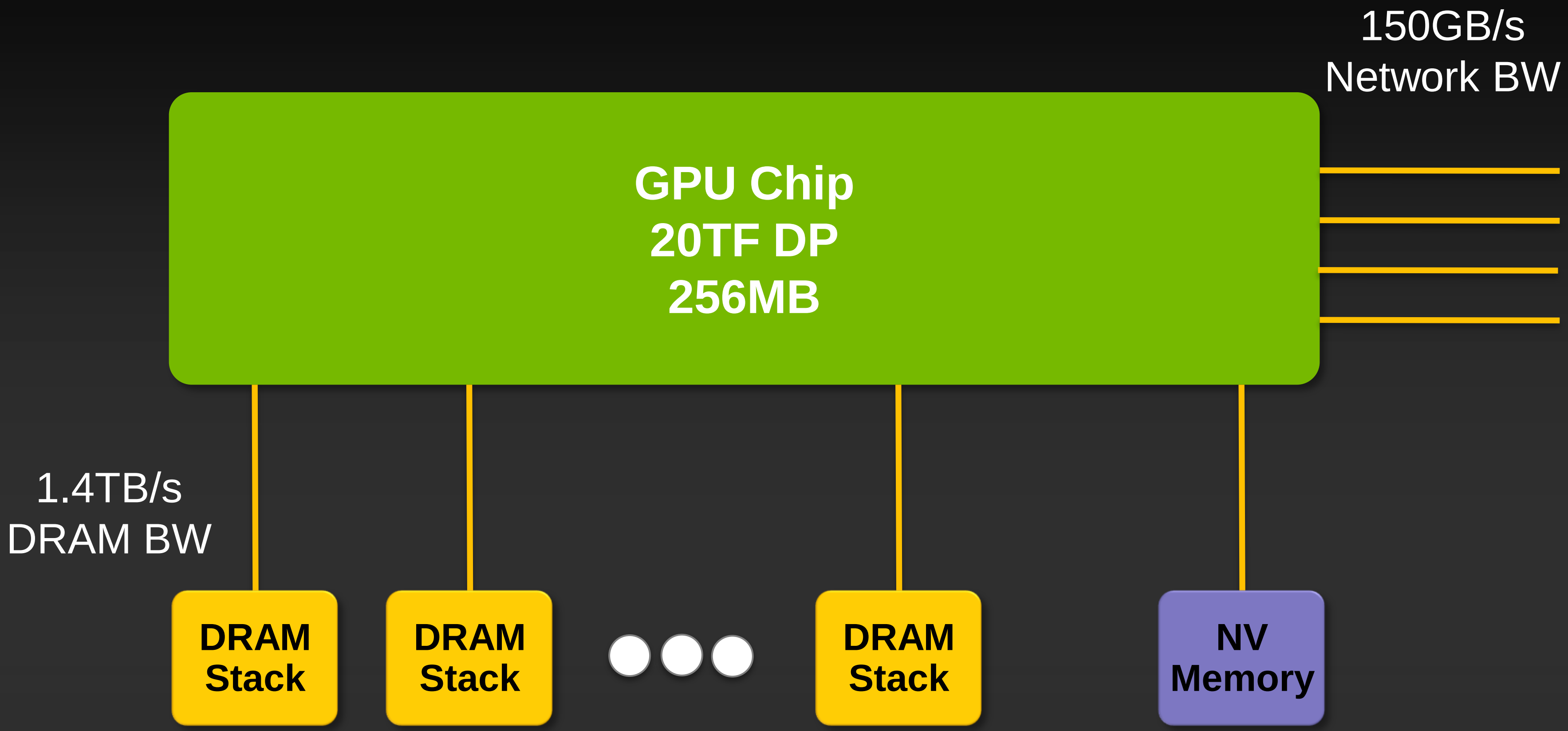
SM – 8 lanes – 160GFLOPS



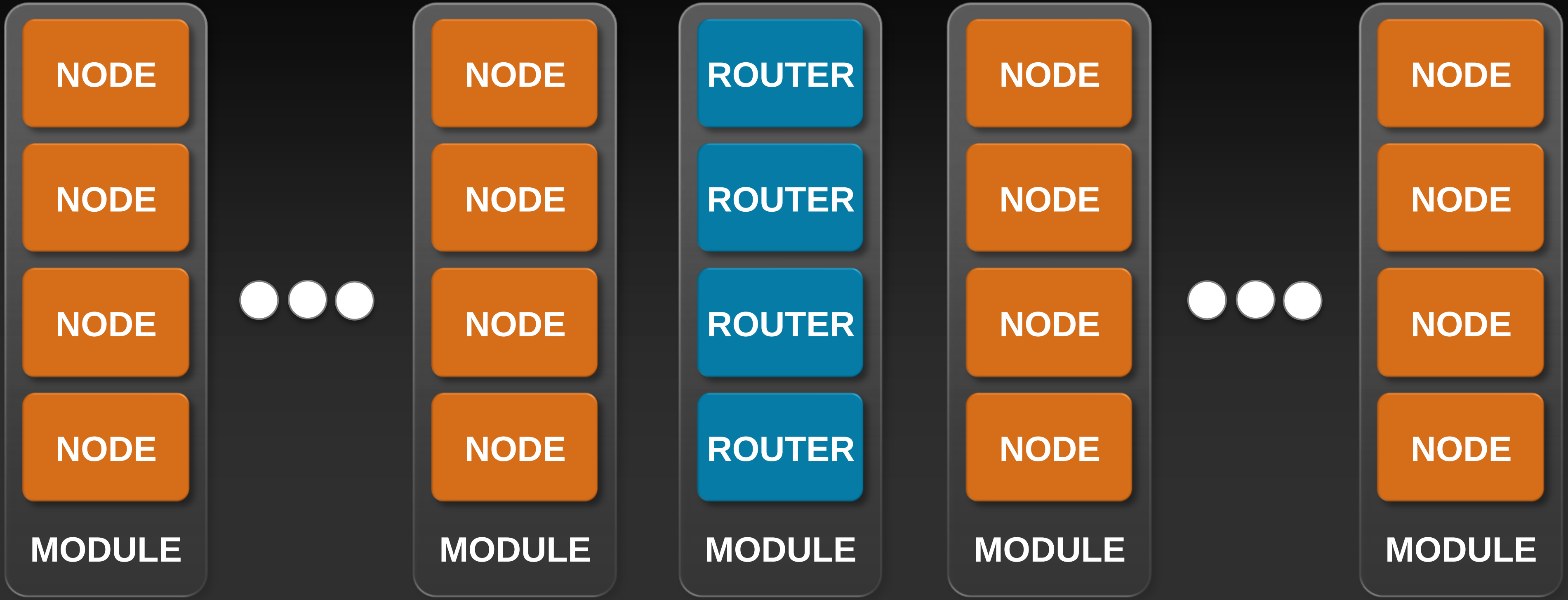
Chip – 128 SMs – 20.48 TFLOPS + 8 Latency Processors



Node MCM – 20TF + 256GB

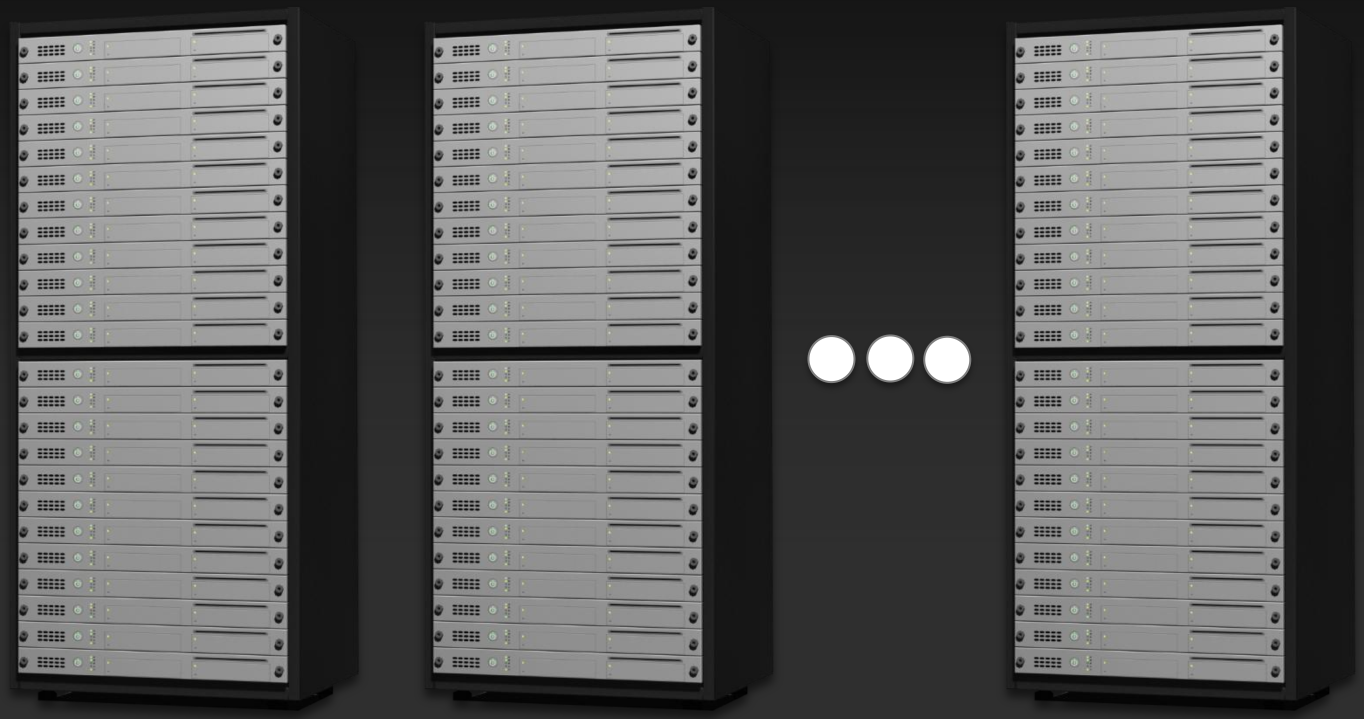


Cabinet – 128 Nodes – 2.56PF – 38 kW



32 Modules, 4 Nodes/Module,
Central Router Module(s), Dragonfly Interconnect

System – to ExaScale and Beyond



Dragonfly Interconnect
400 Cabinets is ~1EF and ~15MW

CONCLUSION

GPU Computing is the Future

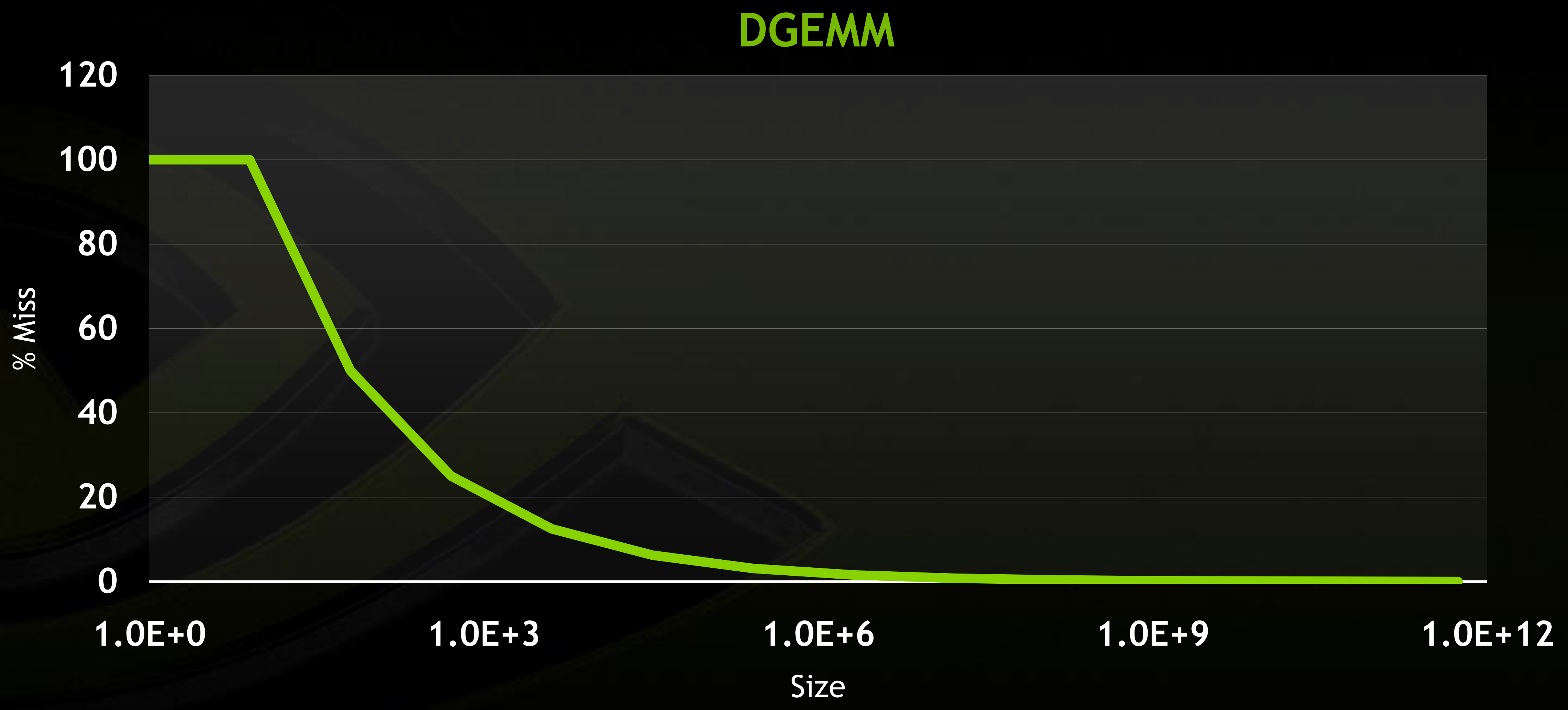
1. **GPU Computing is #1 Today**
   On Top 500 AND Dominant on Green 500
2. **GPU Computing Enables ExaScale**
   At Reasonable Power
3. **The GPU is the Computer**
   A general purpose computing engine, not just an accelerator
4. **The Real Challenge is Software**

THANK YOU

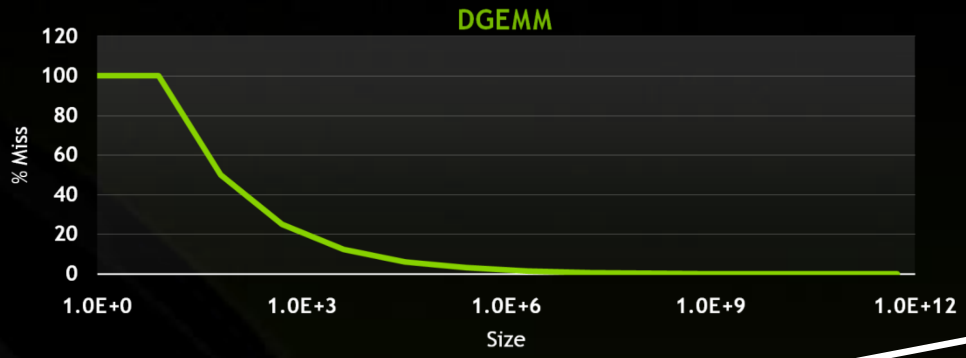
Power is THE Problem

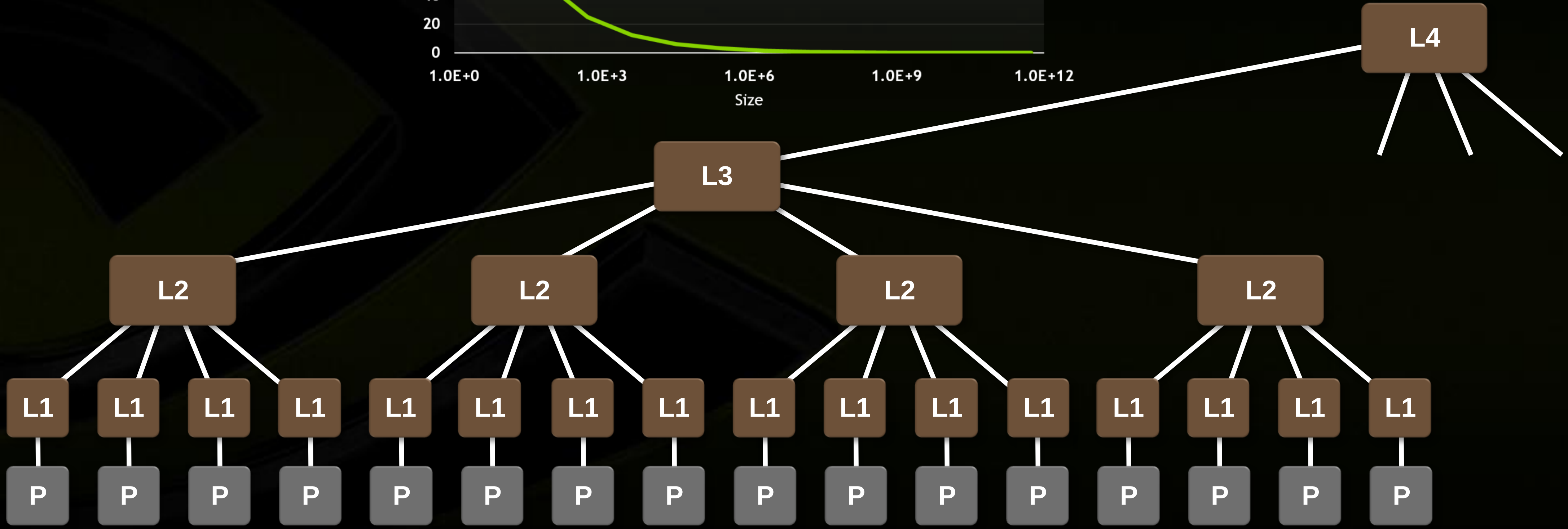
1. Data Movement Dominates Power
2. Optimize the Storage Hierarchy
3. Tailor Memory to the Application

Some Applications Have Hierarchical Re-Use

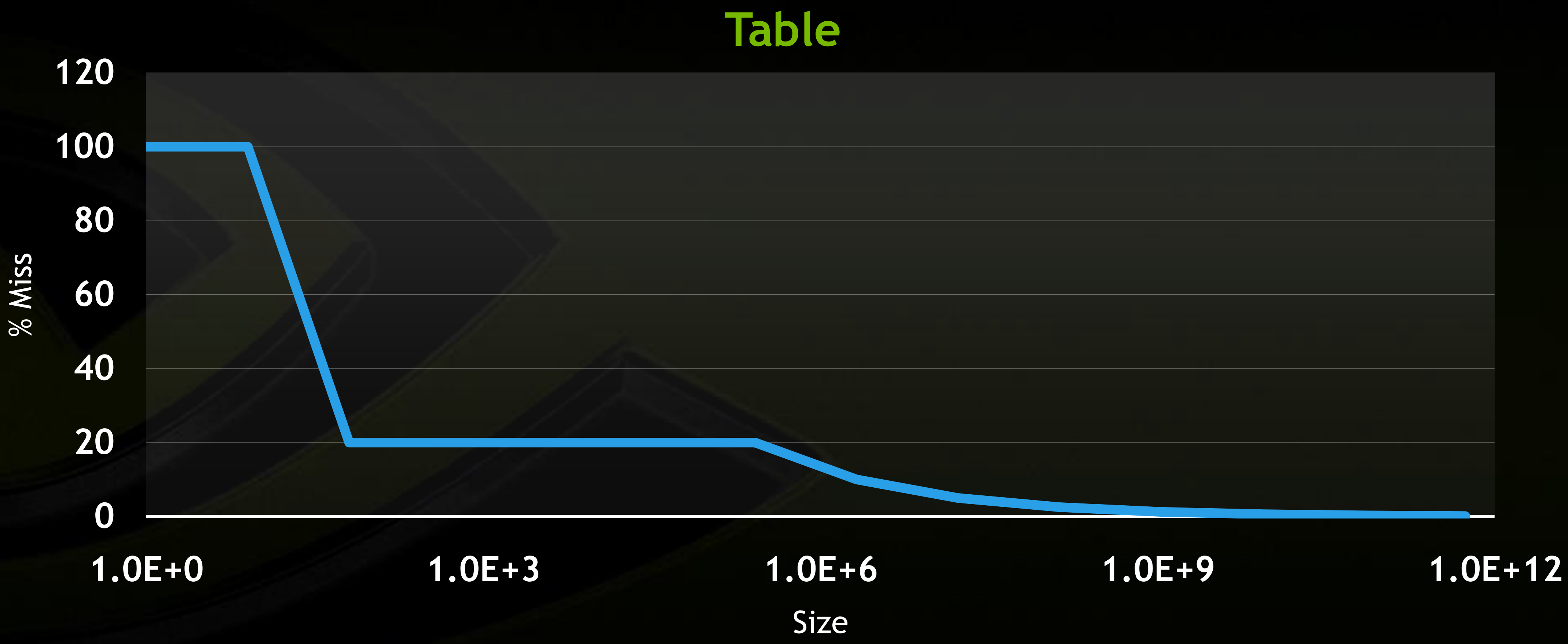


Applications with Hierarchical Reuse Want a Deep Storage Hierarchy

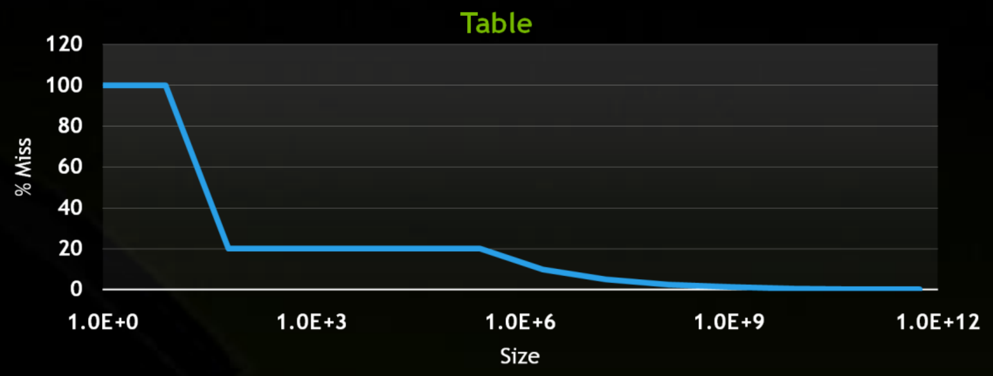


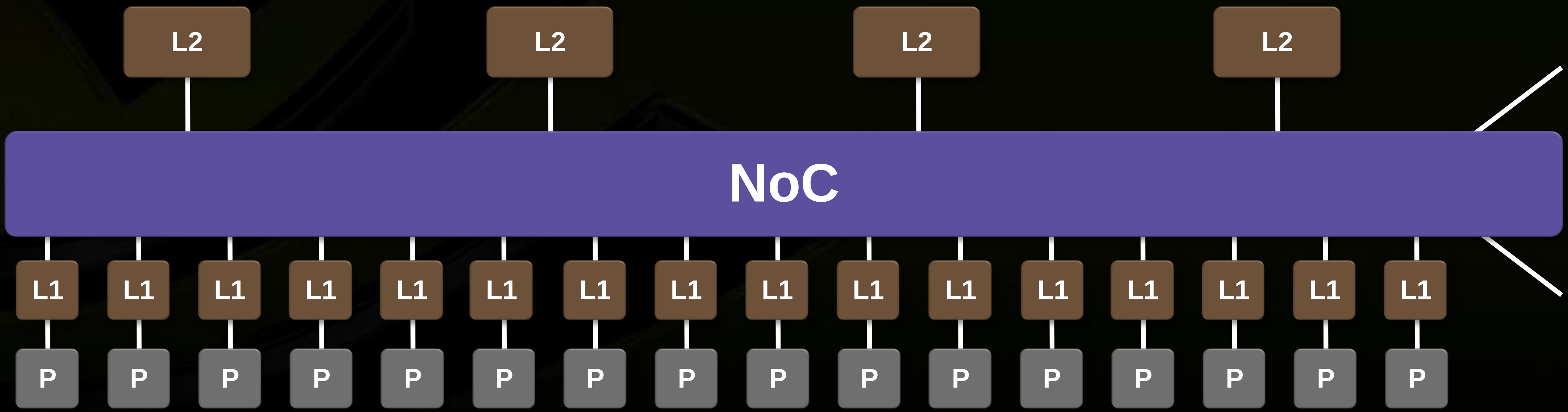


Some Applications Have Plateaus in Their Working Sets



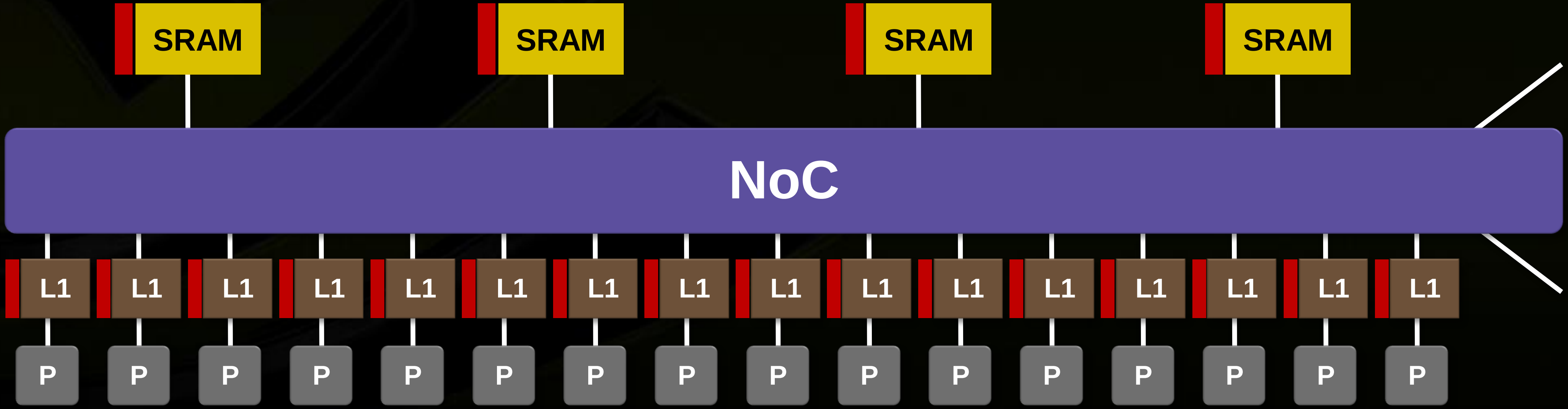
Applications with Plateaus Want a Shallow Storage Hierarchy





Configurable Memory Can Do Both At the Same Time

- Flat hierarchy for large working sets
- Deep hierarchy for reuse
- “Shared” memory for explicit management
- Cache memory for unpredictable sharing



Configurable Memory Reduces Distance and Energy