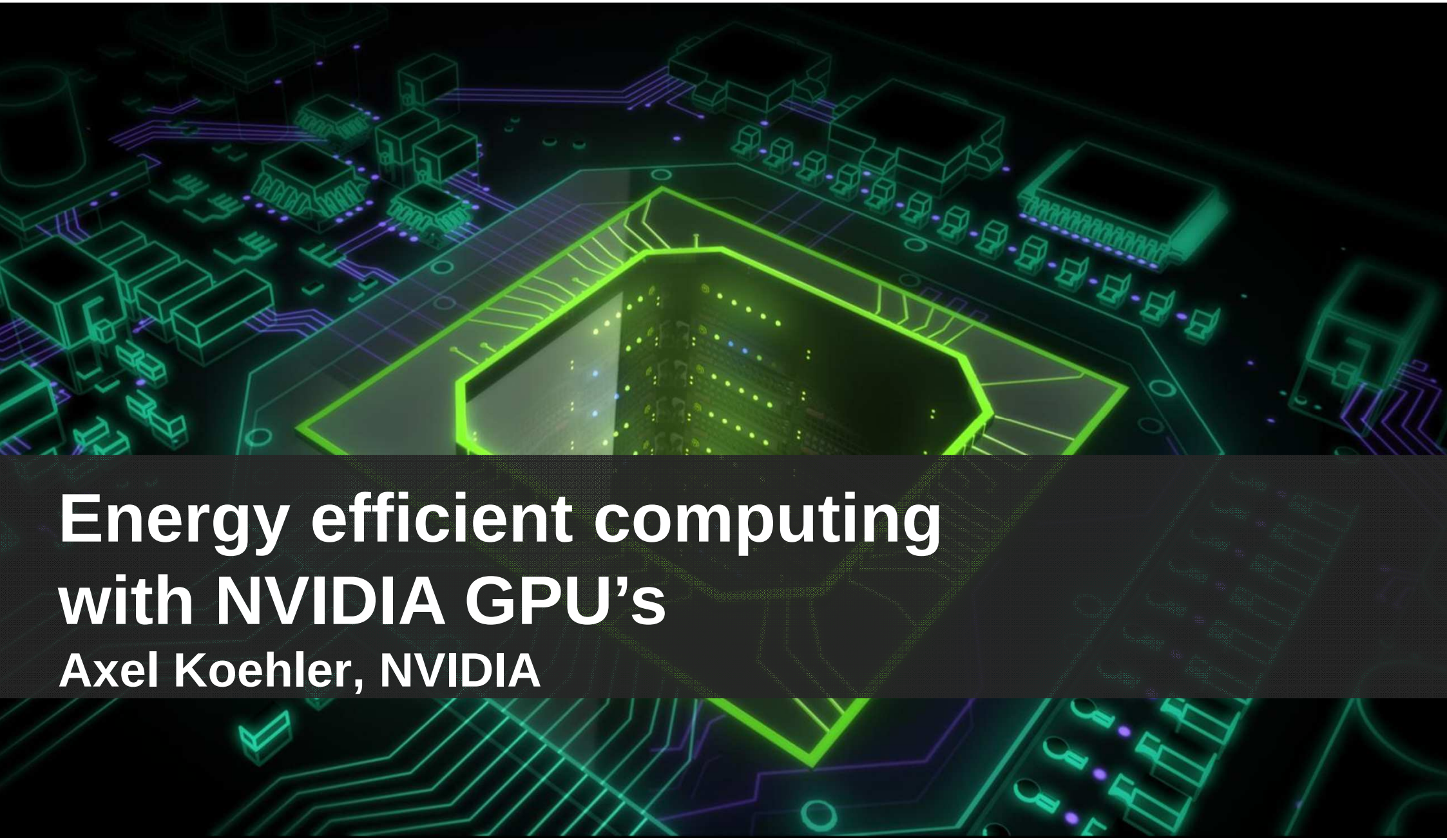
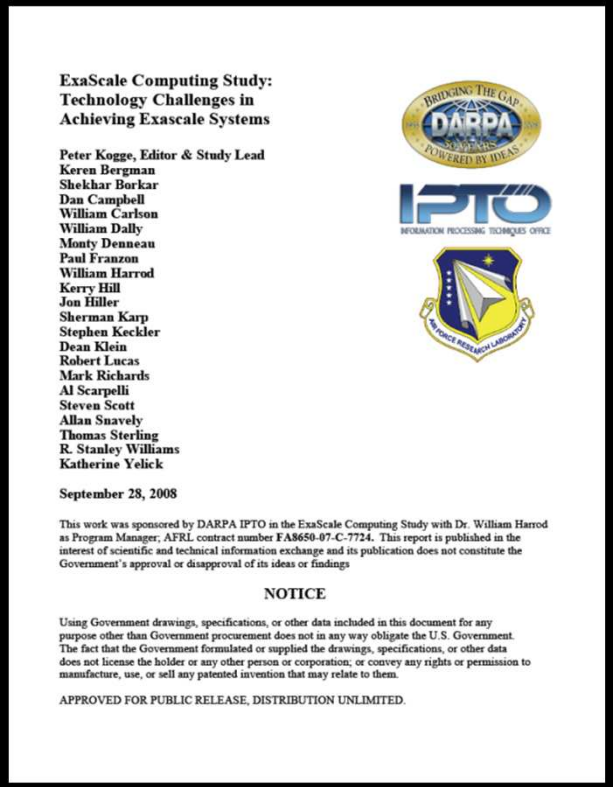


Energy efficient computing with NVIDIA GPU's

Axel Koehler, NVIDIA

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

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

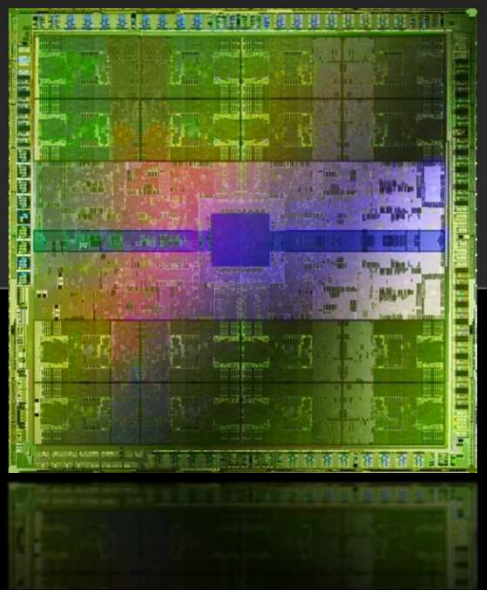
Report published September 28, 2008:

- **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

GPU
200pJ/Instruction

Optimized for Throughput

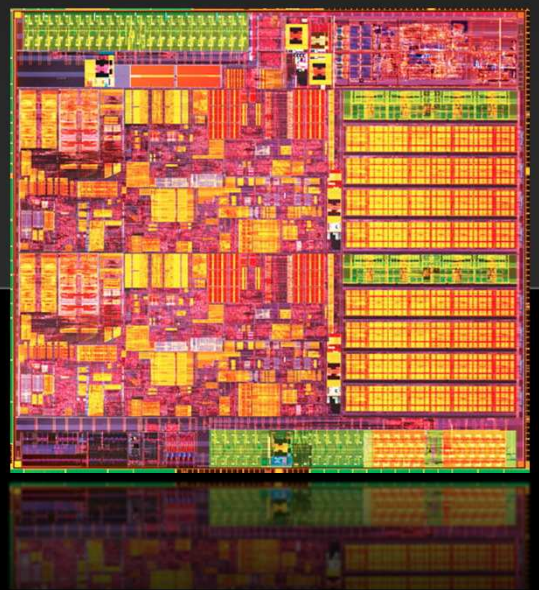
Explicit Management of On-chip Memory



CPU
2nJ/Instruction

Optimized for Latency

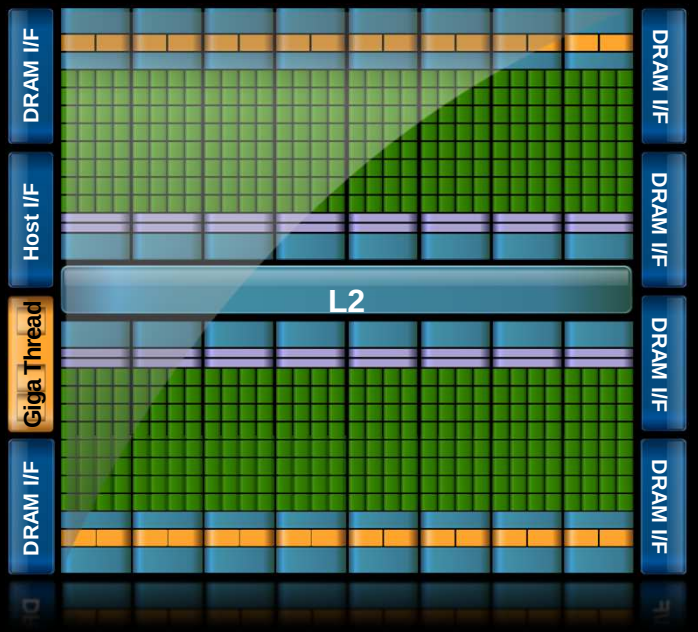
Caches



Energy efficient GPU

Performance = Throughput

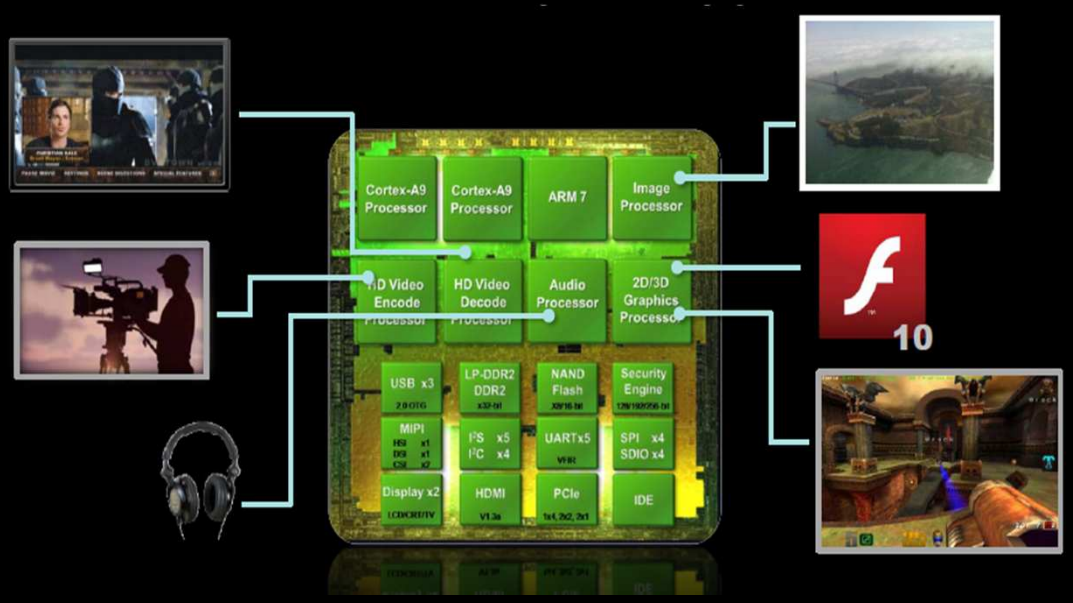
- Fixed function hardware
  - Transistors are primarily devoted to data processing
  - Less leaky cache
- SIMT thread execution
  - Groups of threads formed into warps which always executing same instruction
  - Some threads become inactive when code path diverges
- Cooperative sharing of units with SIMT
  - eg. fetch instruction on behalf of several threads or read memory location and broadcast to several registers
- Lack of speculation reduces overhead
- Minimal Overhead
  - Hardware managed parallel thread execution and handling of divergence



NVIDIA Tegra

- Embedded (Smartphone/Tablet) market is driving energy conservative design
- Aligned requirements for embedded market and HPC

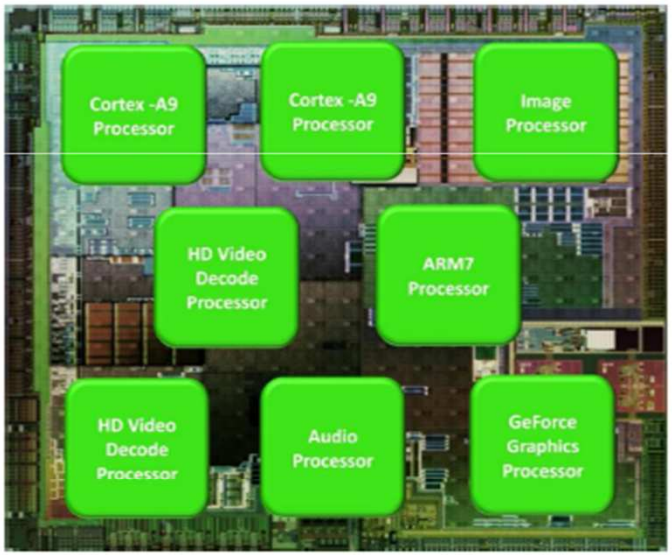
Tegra processor
- Advanced, mobile System-on-a-Chip (SoC)
- Low-power, high performance
- Integrated ARM CPU and GPU



BSC: Ultra-Low Power Clusters using Tegra2 (2/3)

Prototype Description

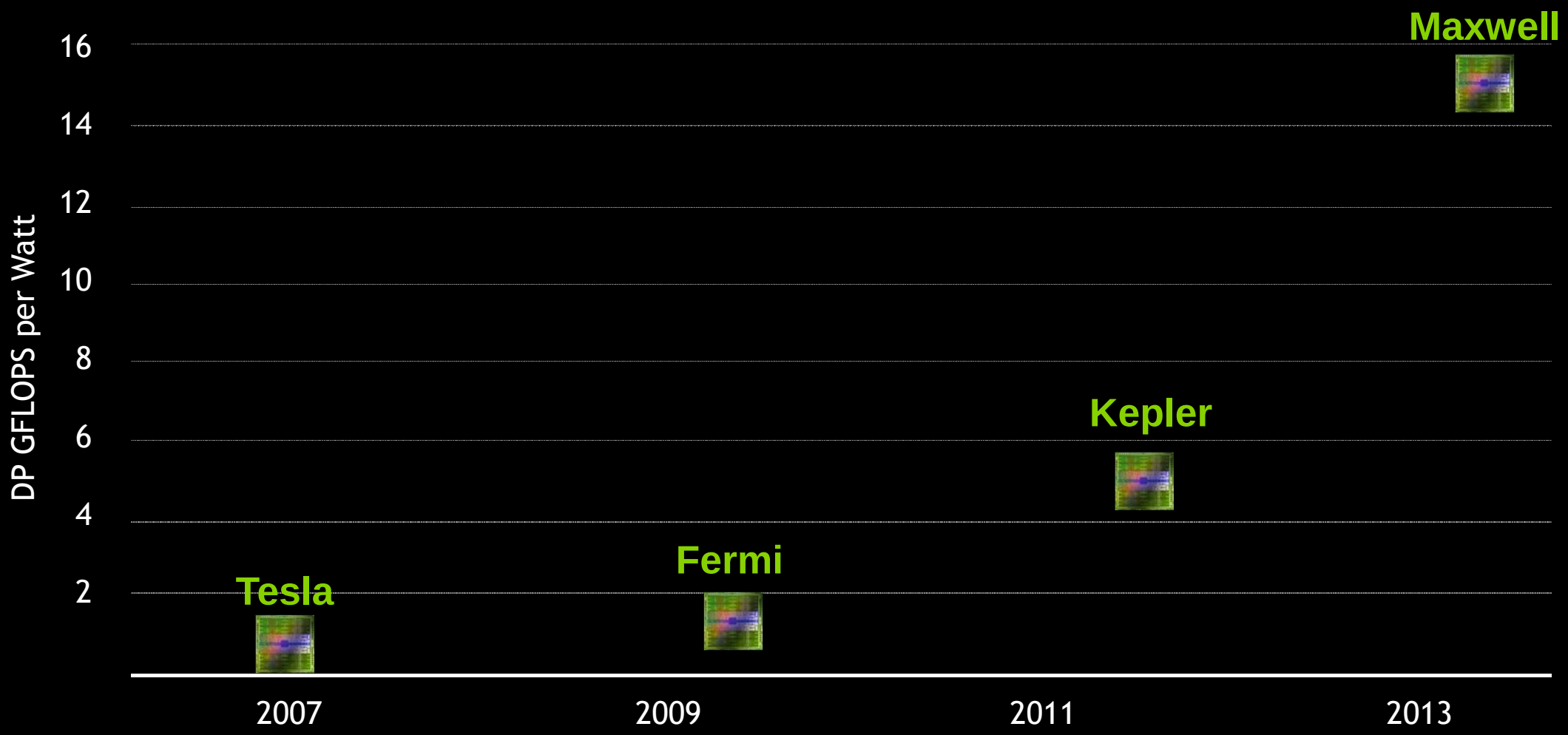
- 128 (or may be 256) NVIDIA Tegra2 boards, each with a Tegra 250 chip.
- Development Board Specs:
  - Tegra250 SoC
    - Dual-core Cortex-A9 @ 1 GHz
  - Ultra low-power GPU
    - OpenGL support only
    - No CUDA or OpenCL
  - 1GB DDR2-667
  - 100Mbit Ethernet
  - 2 x Mini PCIe slots
- Cortex-A9 includes VFP11v3
  - Double-precision
  - Fused Multiply-Add
  - Up to 2 GFLOPS / core
    - 1-2 FP ops / cycle @ 1 GHz
  - 250 mW per core
- 2 GFLOPS / 0.5 Watts
  - ~ 4 GFLOPS/Watt
- Power Management by ARM7 processor
- Support dynamic voltage and frequency scaling



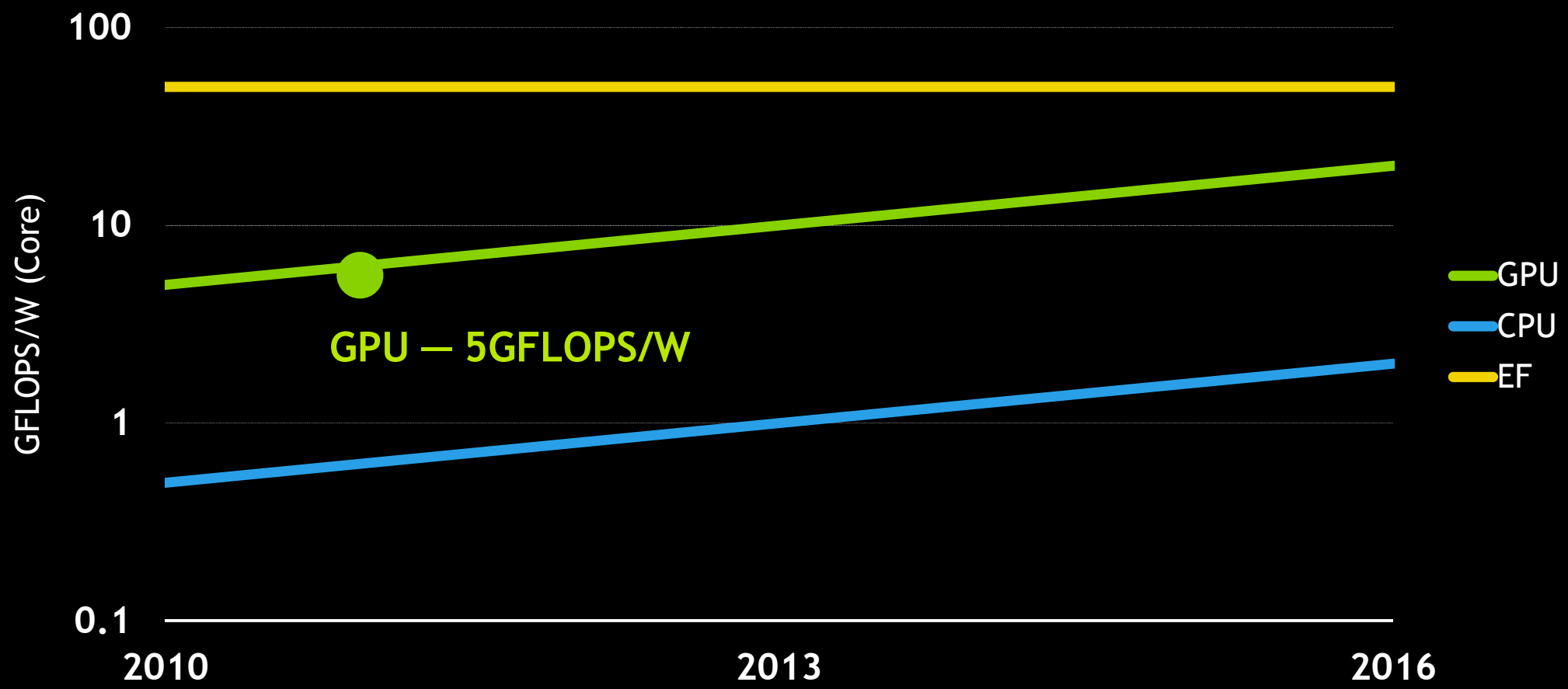
A Tegra250 chip

Courtesy : Tegra whitepaper

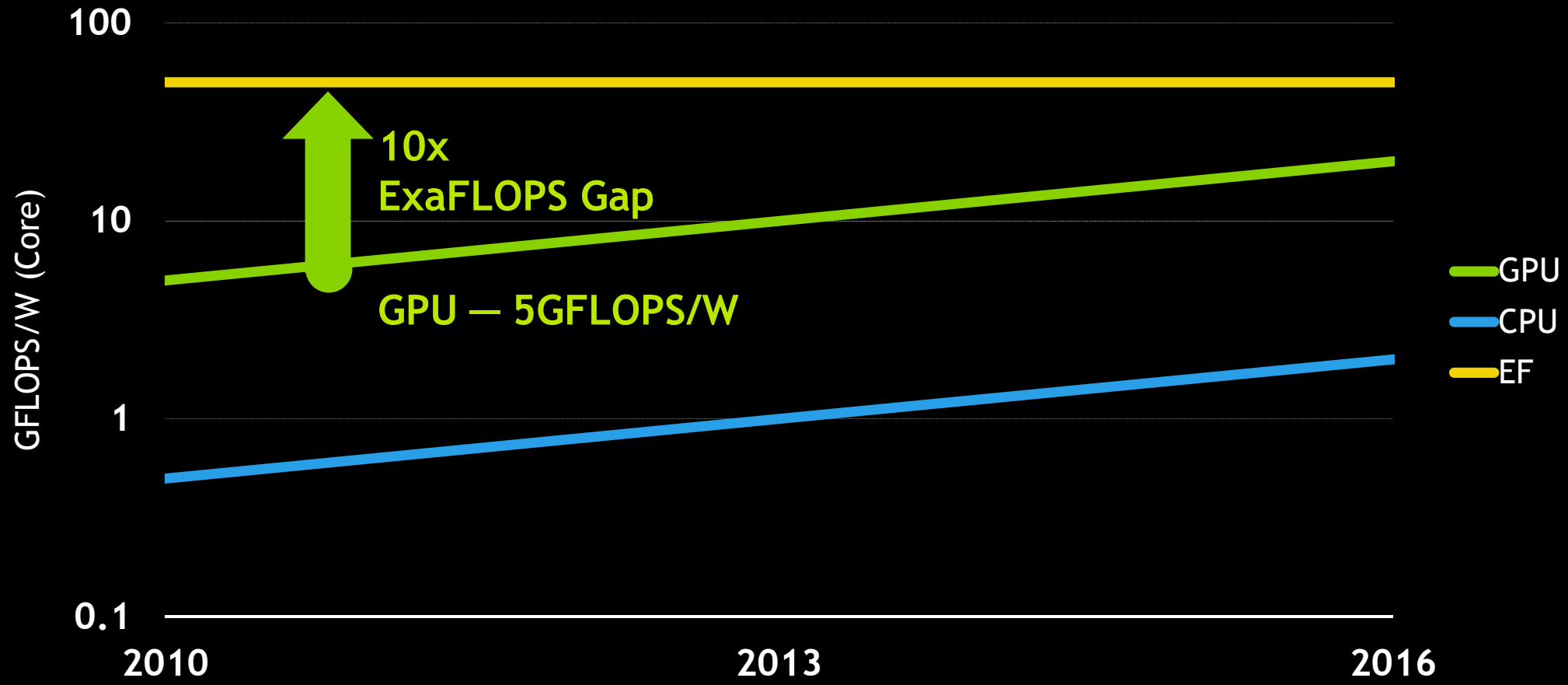
NVIDIA GPU Roadmap



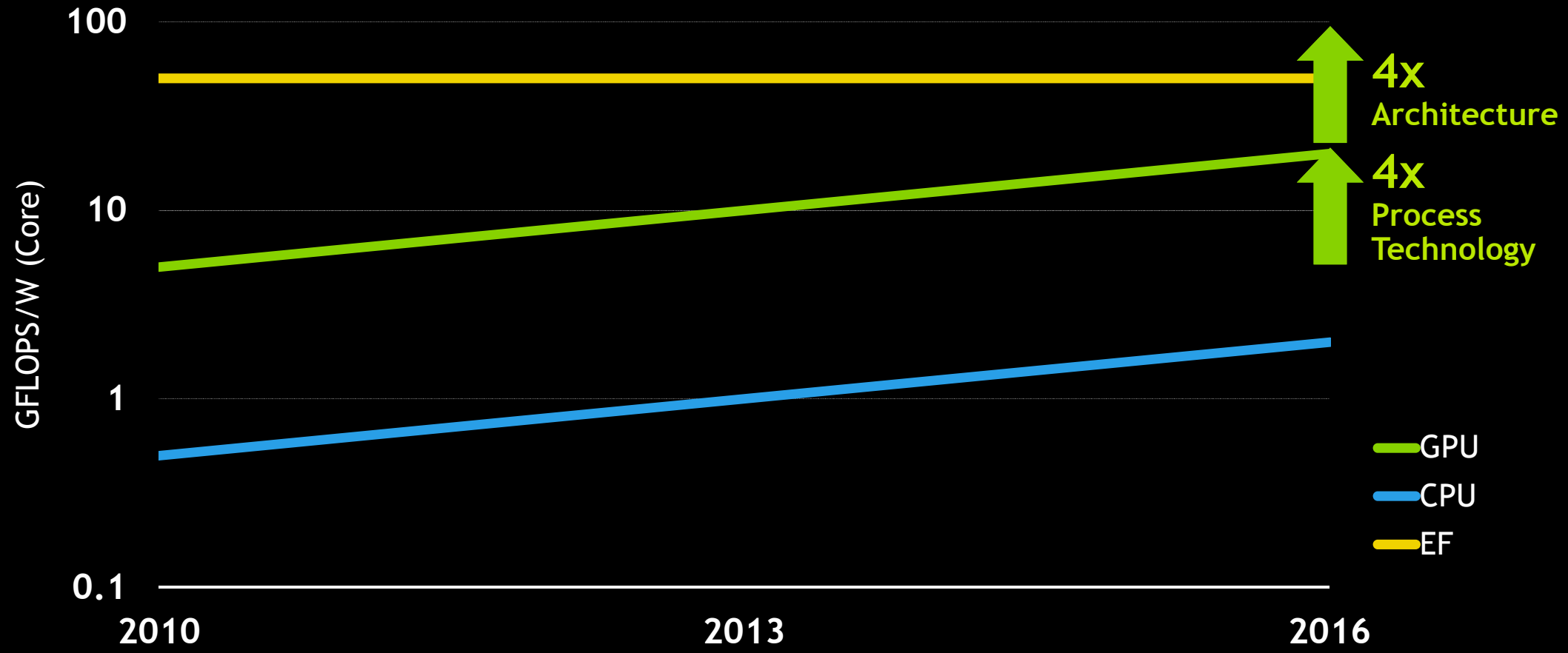
ExaFLOPS at 20MW = 50GFLOPS/W



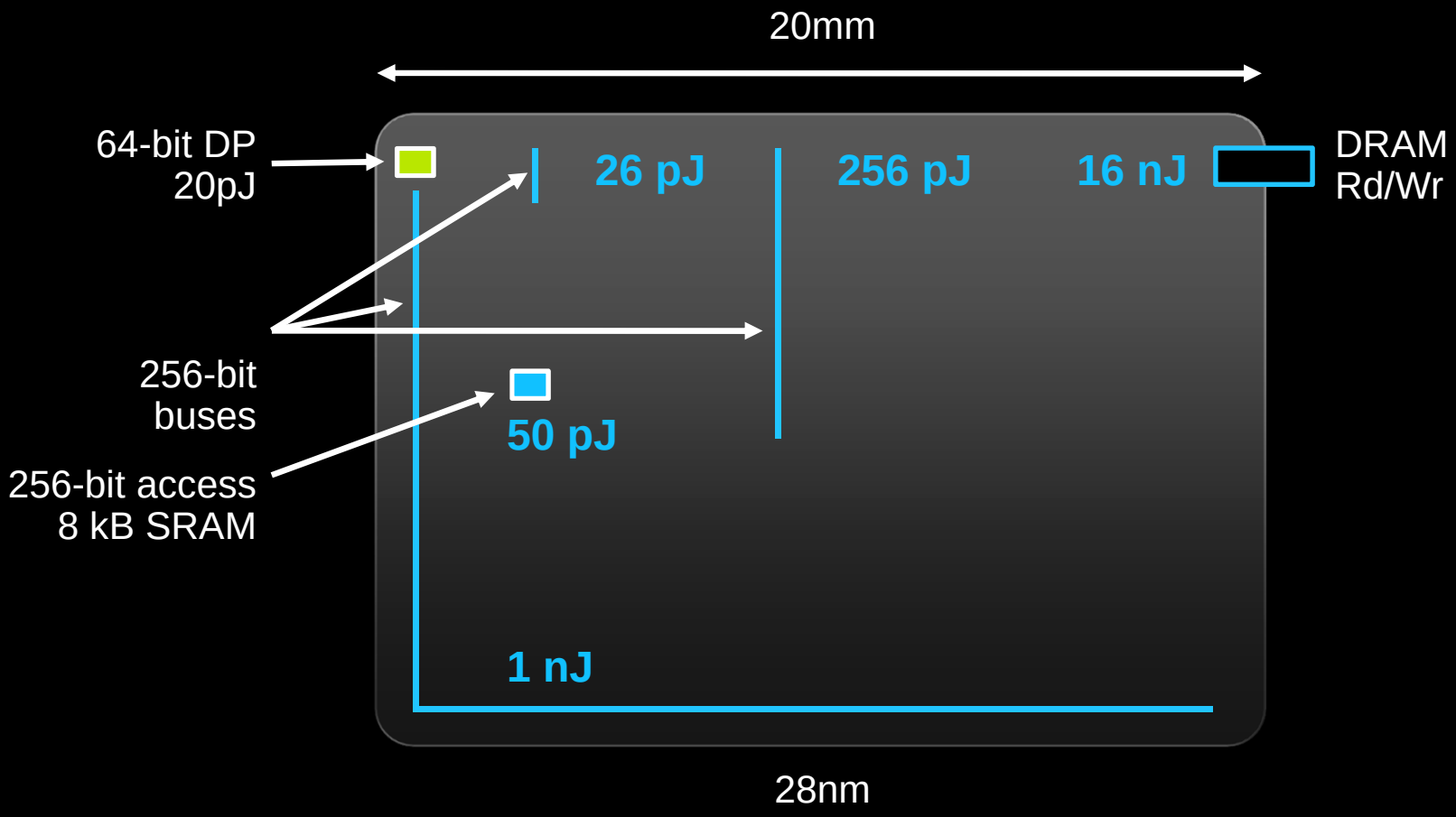
50GFLOPS/W
10x Energy Gap for Today's GPU



GPUs Close the Gap with Process and Architecture



Power is the problem



Fetching operands costs more than computing on them

What is important for the future?

- Its not about the FLOPS
- Its about data movements
- Algorithms should be designed to perform more work per unit data movement
- Programming systems should further optimize this data movement
- Architectures should facilitate this by providing an exposed hierarchy and efficient communication

Addressing The Power Challenge
Locality and Overhead

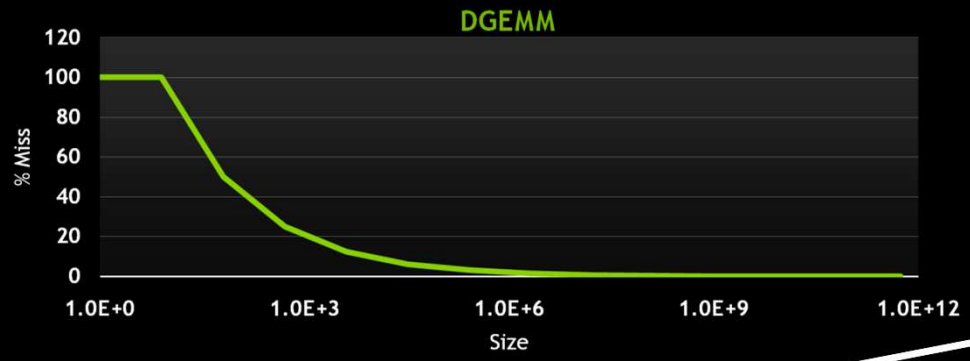
- Locality
  - Bulk of data must be accessed from nearby memories (2pJ) not across the chip (150pJ), off chip (300pJ) or across the system (1nJ)
  - Application, programming system, and architecture must work together to exploit locality
- Overhead
  - Bulk of execution energy must go to carrying out the operation not scheduling instructions (100x today)
  - We must build efficient cores – where the bulk of the energy is spent on operations, not overhead
    - An Out-of-Order Core spends 2nJ to schedule a 25pJ FMUL (or an 0.5pJ integer add)

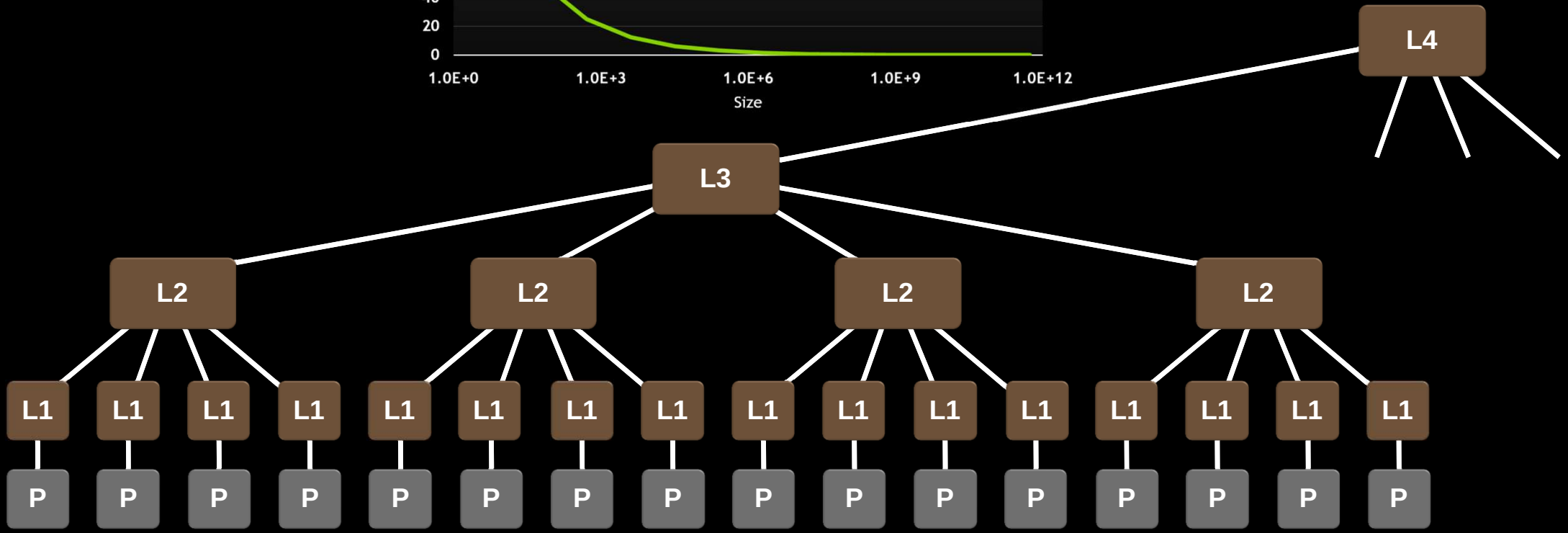
Echelon Team

Echelon project goals

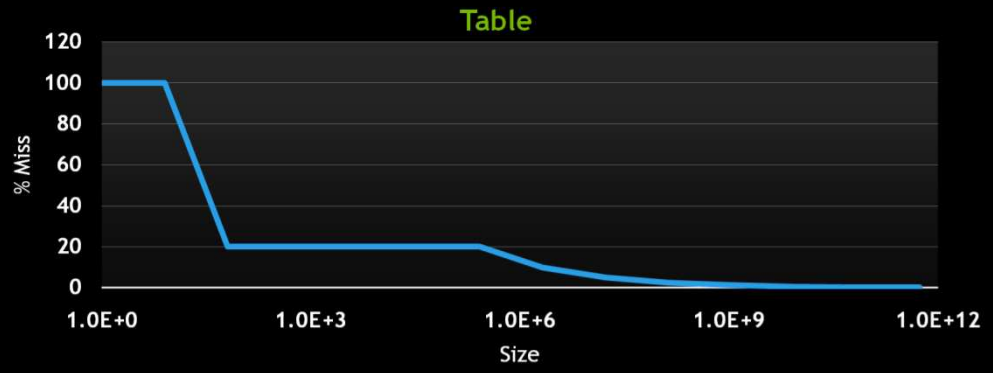
- Develop a design for an ExaScale system
- Optimize data movements which dominates the power
- Optimize the storage hierarchy
- Tailor the memory to the application

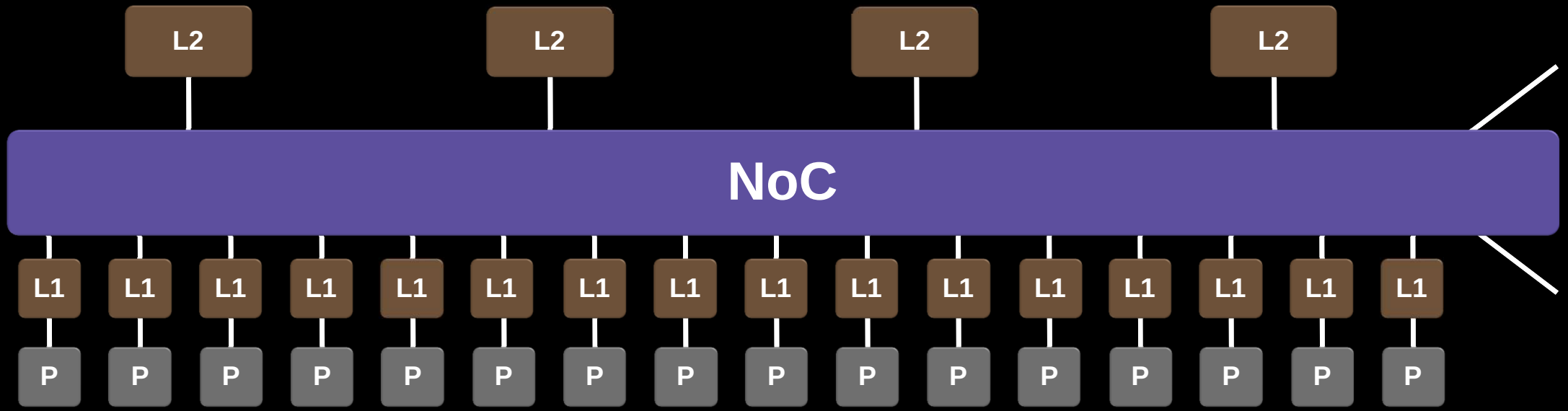
Applications with Hierarchical Reuse Want a Deep Storage Hierarchy





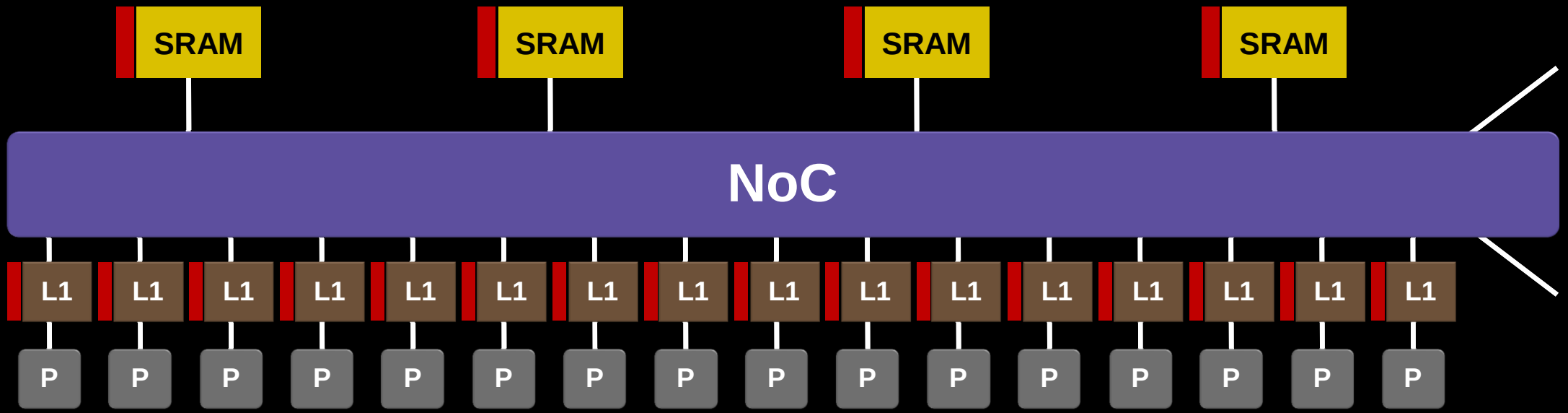
Applications with Plateaus Want a Shallow Storage Hierarchy



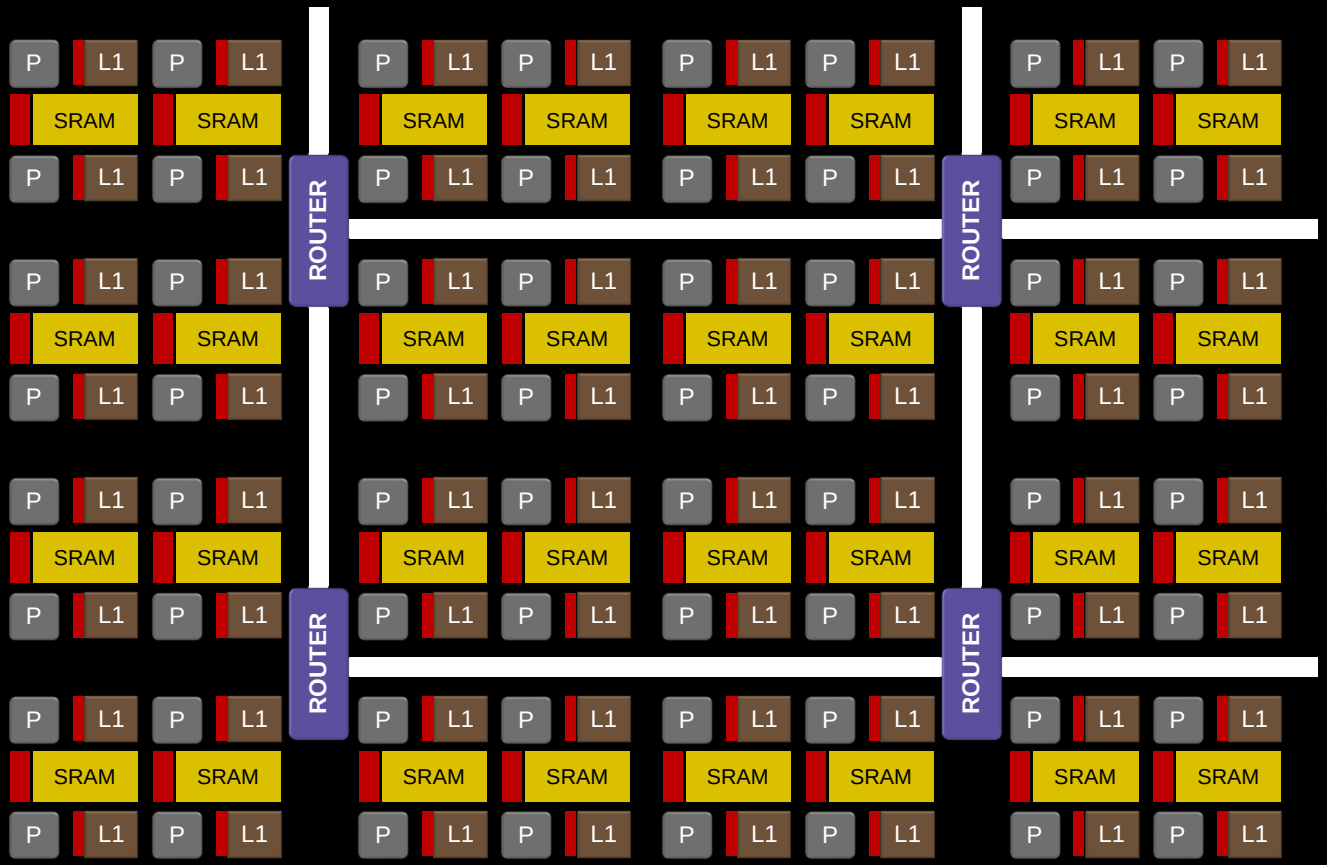


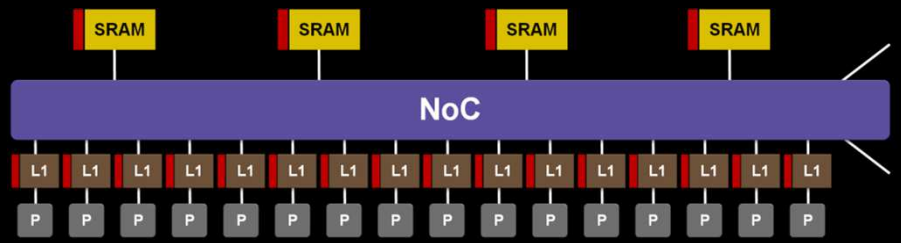
Configurable Memory Can Do Both At the Same Time

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

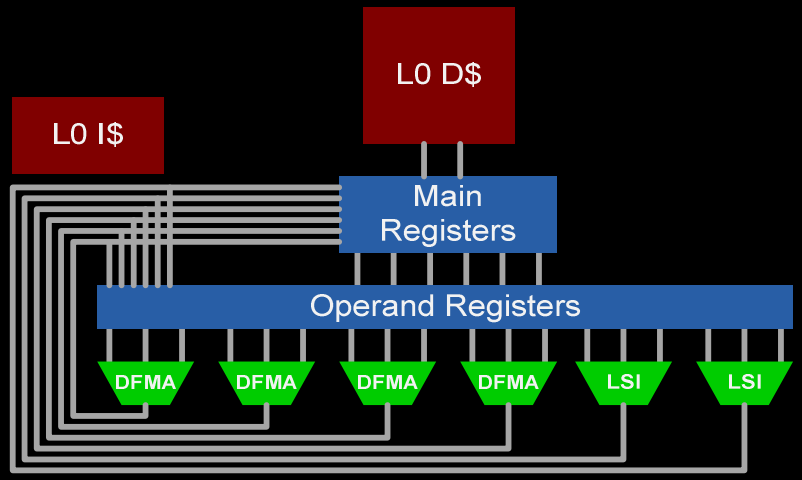


Configurable Memory Reduces Distance and Energy

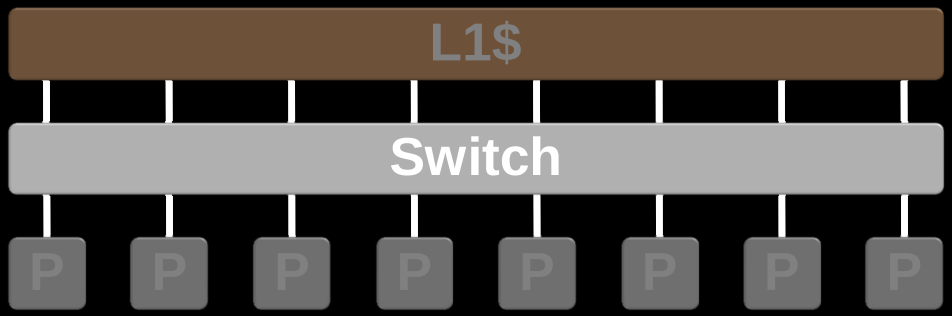




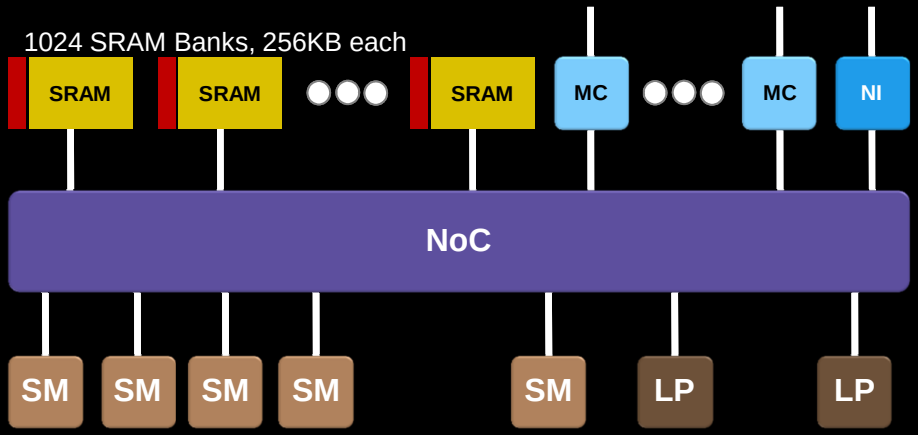
Echelon Architecture (1/2)



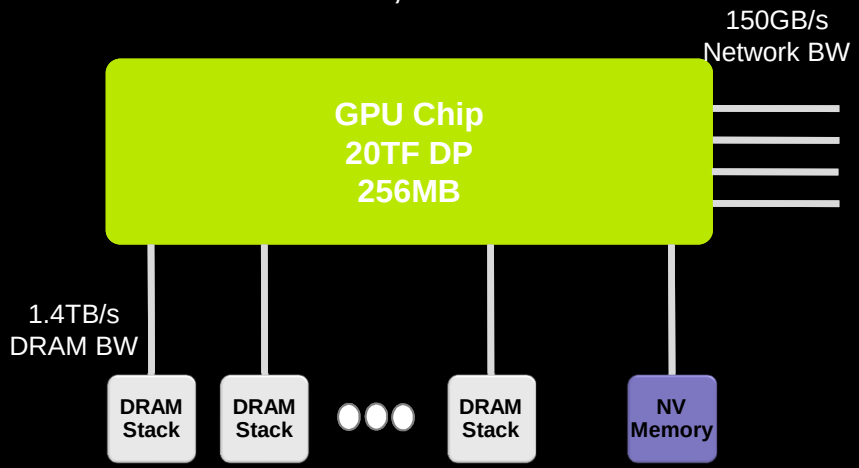
Lane - DFMAs, 20 GFLOPS



SM - 8 lanes, 160 GFLOPS

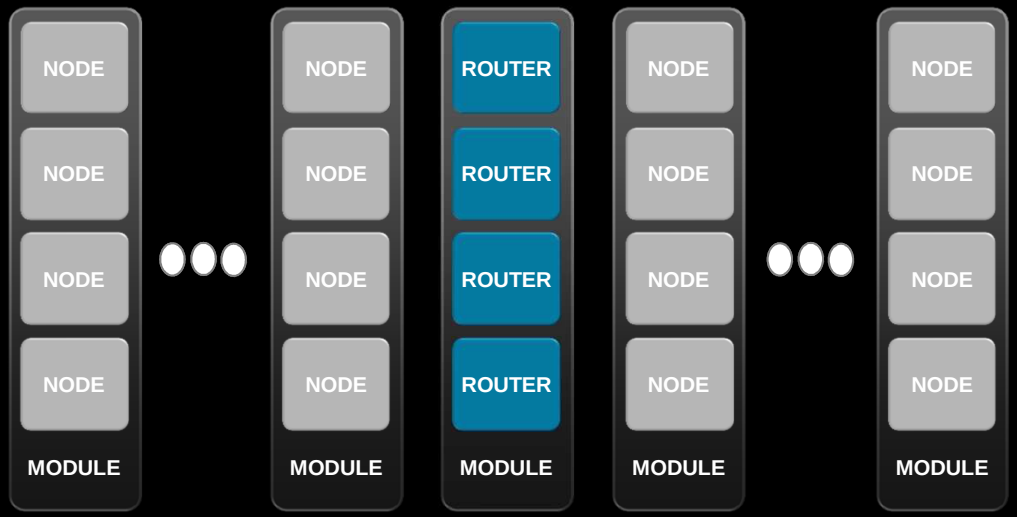


Chip – 128 SMs, 20.48 TFLOPS + 8 Latency Processors



Node MCM – 20 TFLOPS + 256 GB

Echelon Architecture (2/2)

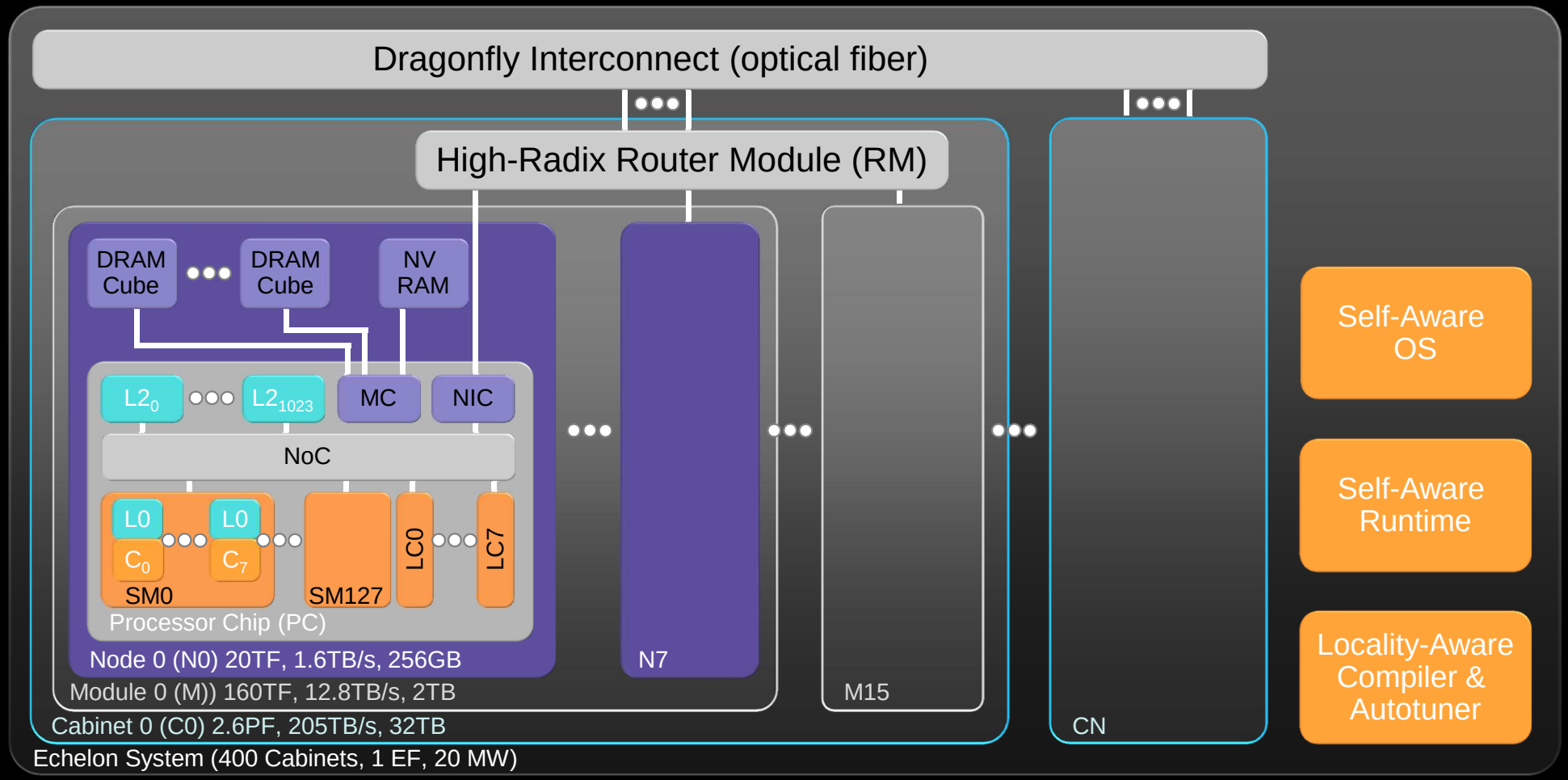


Cabinet – 128 Nodes – 2.56 PF – 50 KW
Central Router Module(s), Dragonfly Interconnect

System – 400 Cabinets – 1 EF – 20 MW
Dragonfly Interconnect

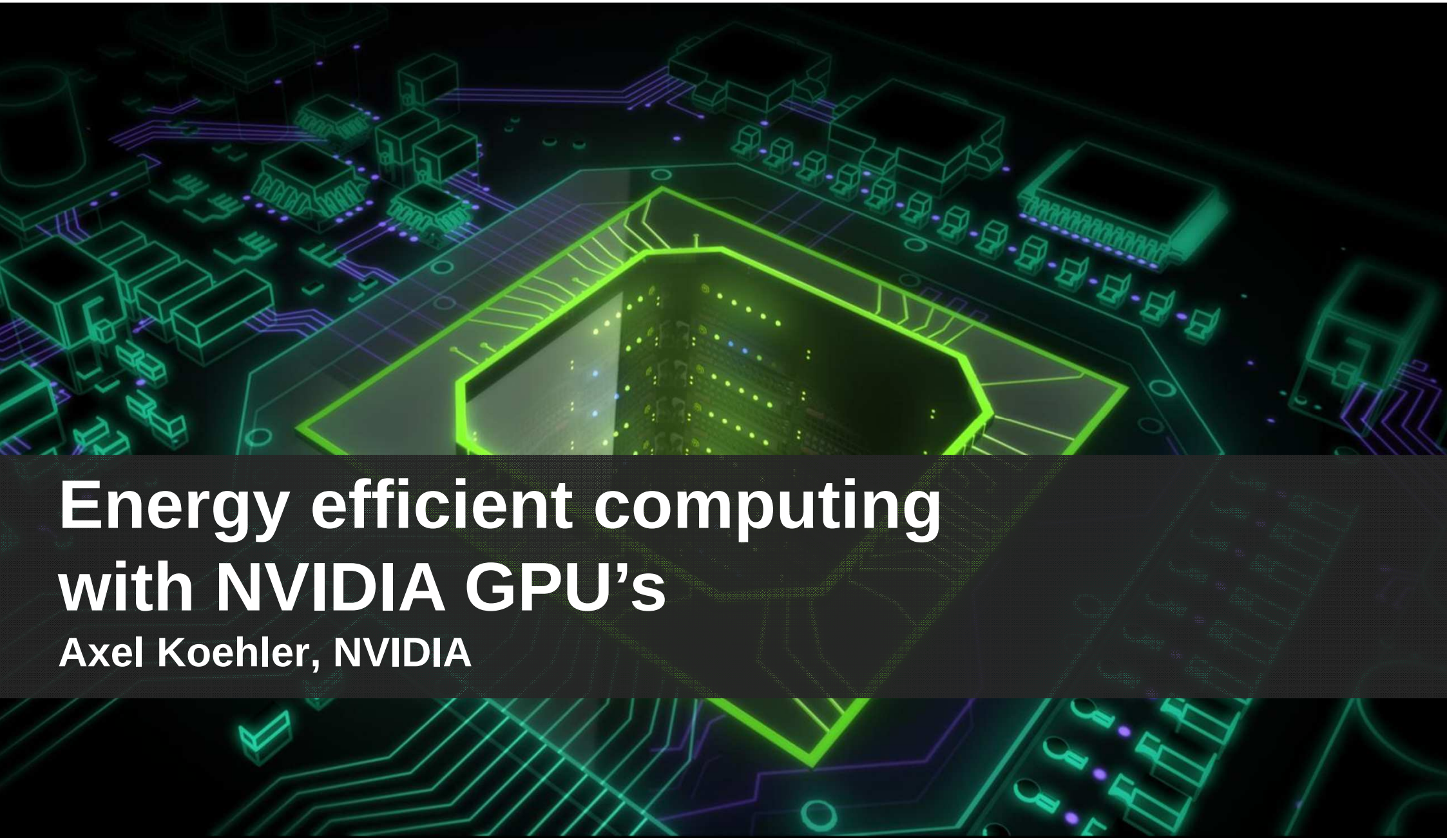
Echelon System Sketch

An Nvidia ExaScale Machine



Summary

- Today's Tesla GPU's are already designed for energy efficiency
- NVIDIA benefits from the experiences in other markets (Tegra)
- Power is the #1 issue for ExaScale Systems
- Data Movement dominates the power
  - Locality at all levels and reduction of overhead is necessary
- Echelon project addresses ExaScale issues



Energy efficient computing with NVIDIA GPU's

Axel Koehler, NVIDIA