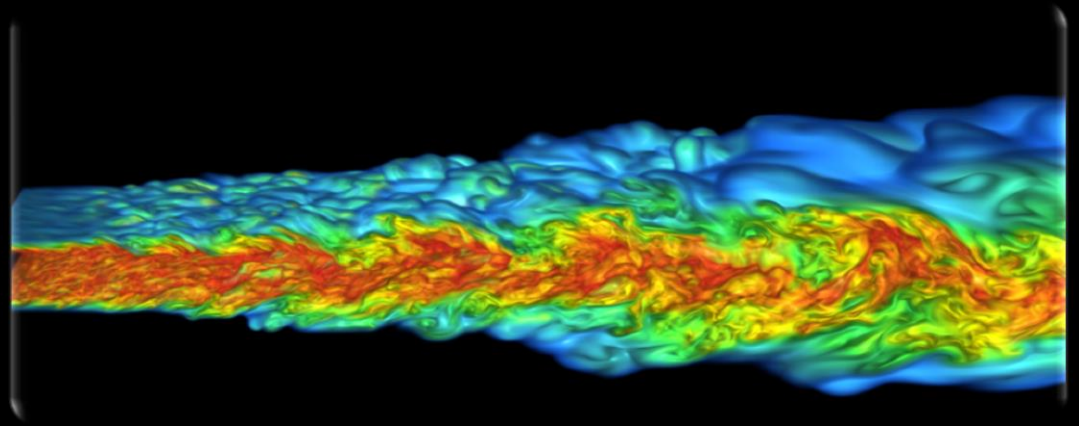
THE PATH TO EXASCALE COMPUTING

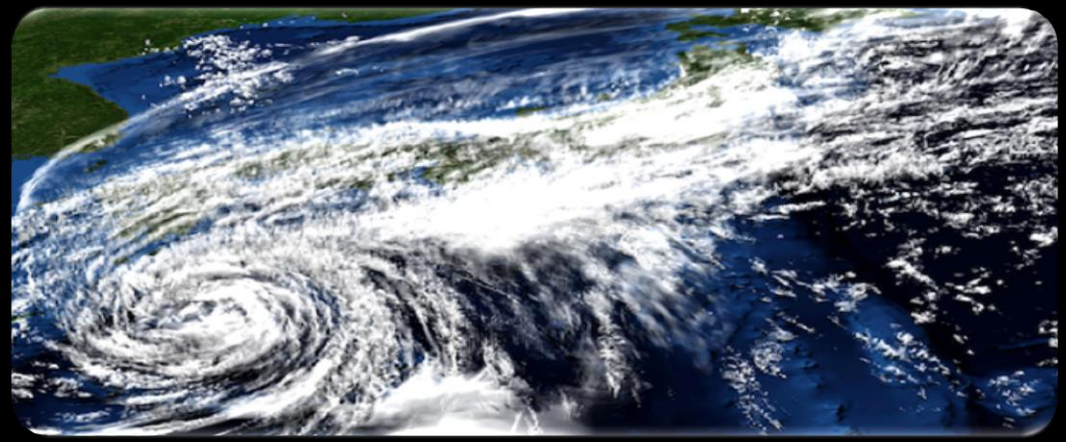
Bill Dally

Chief Scientist and Senior Vice President of Research

The Goal:

Sustained ExaFLOPs on problems of interest





Exascale Challenges

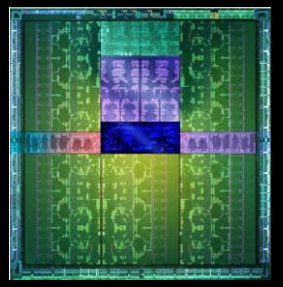
- Energy efficiency
- Programmability
- Resilience
- Sustained performance on real applications
- Scalability

NVIDIA's ExaScale Vision

- Energy efficiency
  - Hybrid architecture, efficient architecture, aggressive circuits, data locality
- Programmability
  - Target-independent programming, adaptation layer, agile network, hardware support
- Resilience
  - Containment domains, low SDC
- Sustained performance on real applications
- Scalability

NVIDIA's ExaScale Vision

- Energy efficiency
  - Hybrid architecture, efficient architecture, aggressive circuits, data locality
- Programmability
  - Target-independent programming, adaptation layer, agile network, hardware support
- Resilience
  - Containment domains, low SDC
- Sustained performance on real applications
- Scalability



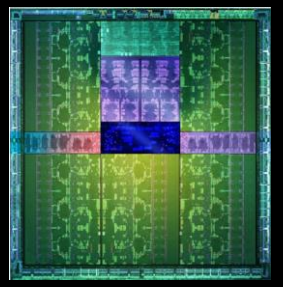
2013

2023

1,000PF (50x)
72,000HCNs (4x)
20MW (2x)
50 GFLOPs/W (25x)
~10^{10} Threads (1000x)

20PF
18,000 GPUs
10MW
2 GFLOPs/W
~10^{7} Threads

You Are Here



2013

20PF
18,000 GPUs
10MW
2 GFLOPs/W
~10^7 Threads

CORAL
150-300PF (5-10x)
11MW (1.1x)
14-27 GFLOPs/W (7-14x)
Lots of Threads

2017

You Are Here

2023

1,000PF (50x)
72,000HCNs (4x)
20MW (2x)
50 GFLOPs/W (25x)
~10^{10} Threads (1000x)

Energy Efficiency

Its not about the FLOPs

16nm chip, 10mm on a side, 200W

A chip with 10^4 FPUs:

$100mm^2$

200W

20TFLOPS

Pack 50,000 of these in racks

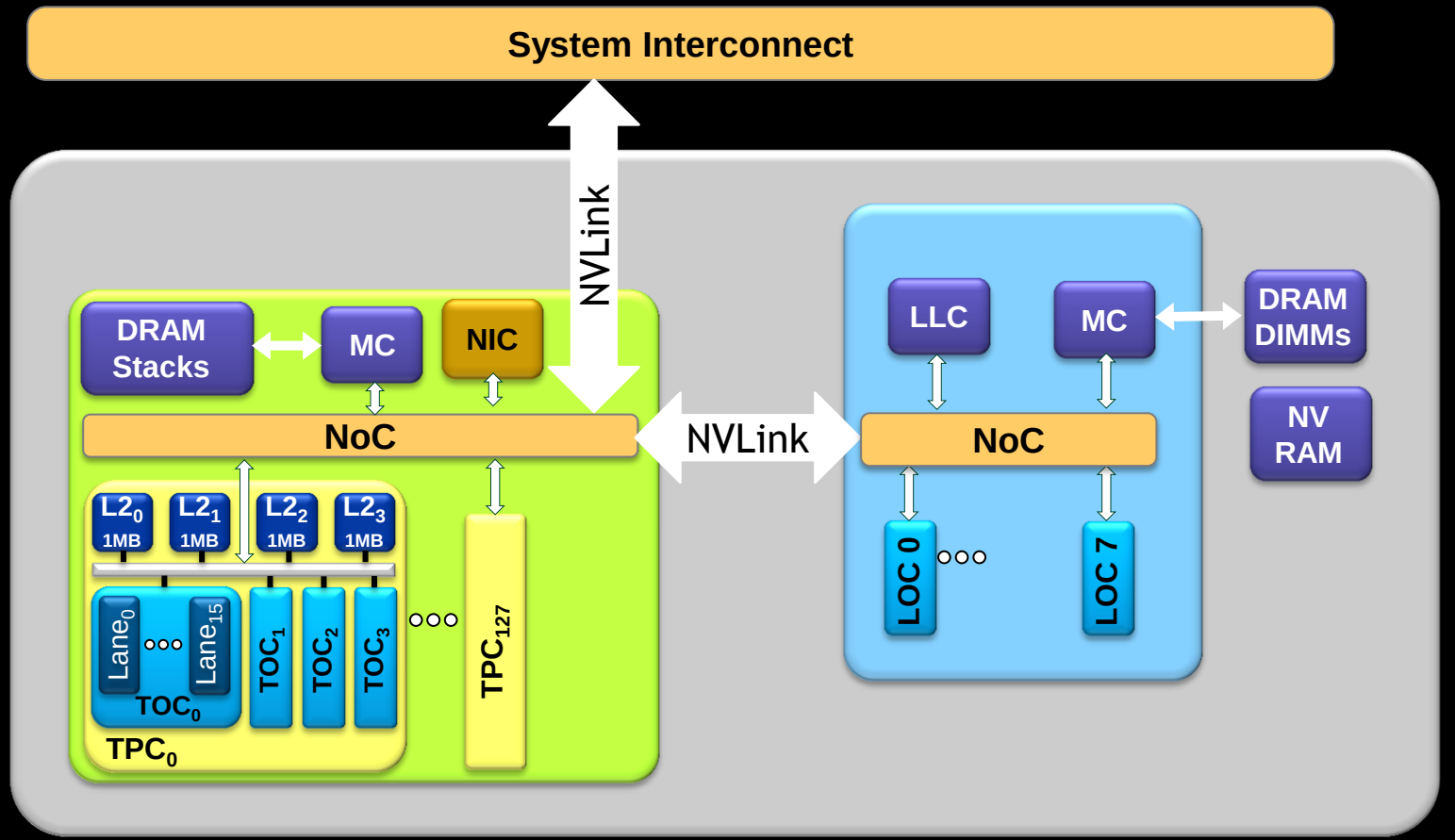
1EFLOPS

10MW

Overhead

Locality

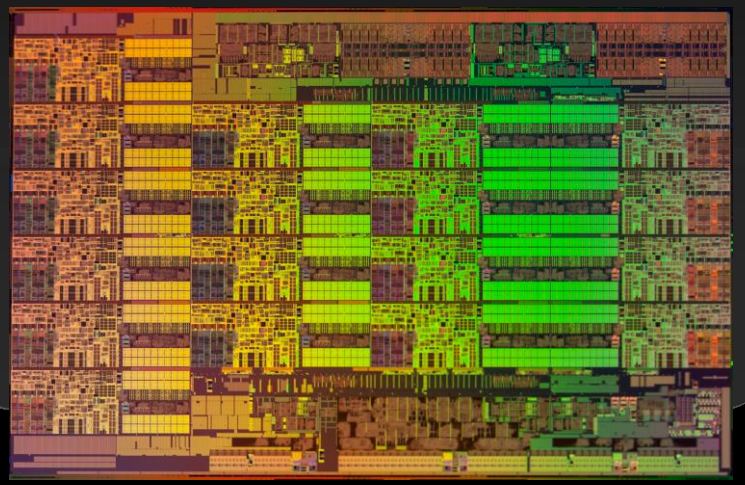
Heterogeneous Node



CPU
130 pJ/flop (Vector SP)

Optimized for Latency

Deep Cache Hierarchy

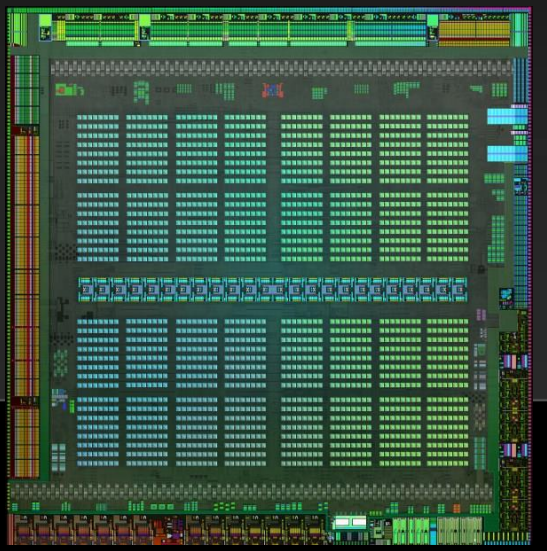


Haswell
22 nm

GPU
30 pJ/flop (SP)

Optimized for Throughput

Explicit Management
of On-chip Memory

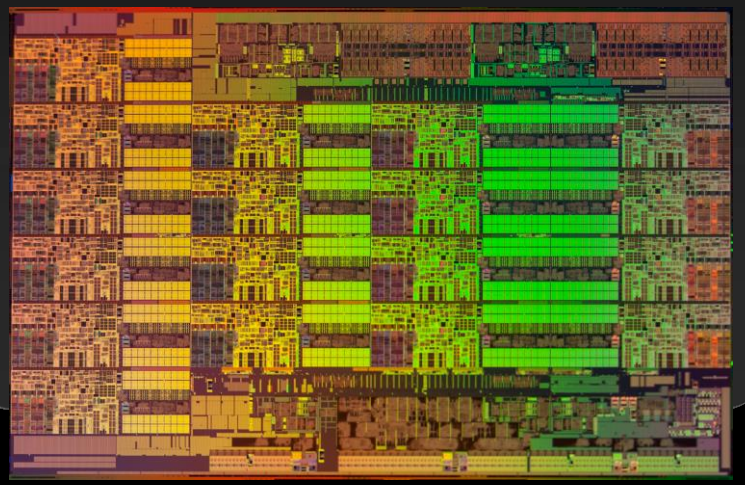


Maxwell
28 nm

CPU
2nJ/flop (Scalar SP)

Optimized for Latency

Deep Cache Hierarchy

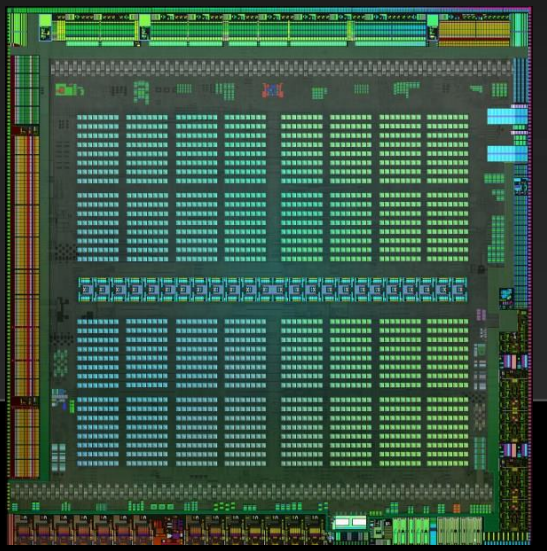


Haswell
22 nm

GPU
30 pJ/flop (SP)

Optimized for Throughput

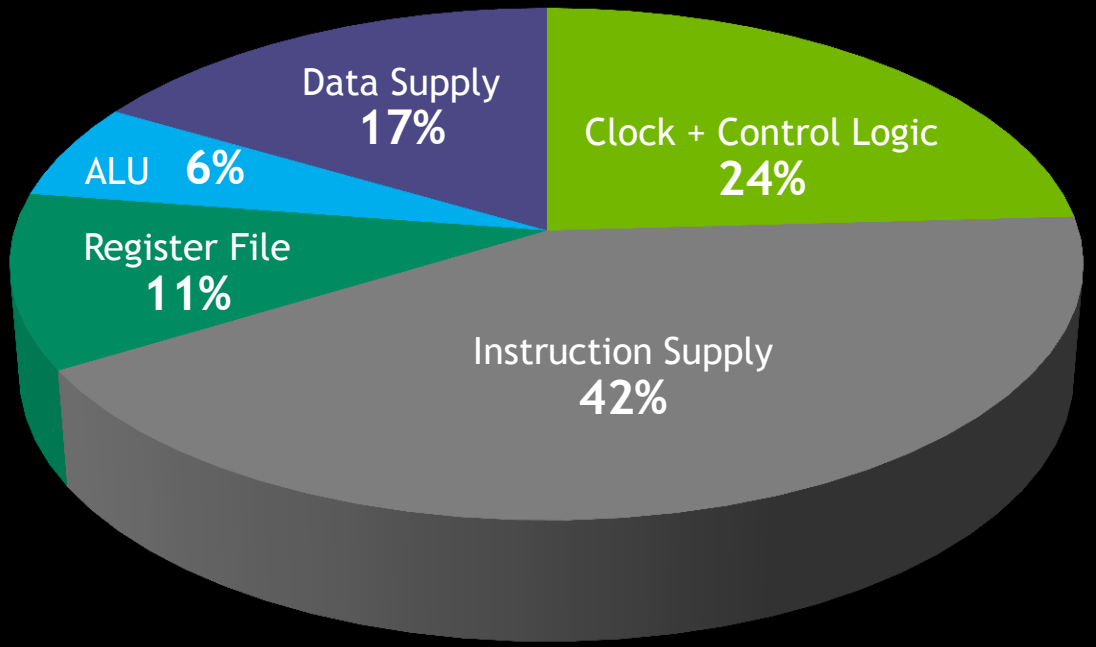
Explicit Management
of On-chip Memory



Maxwell
28 nm

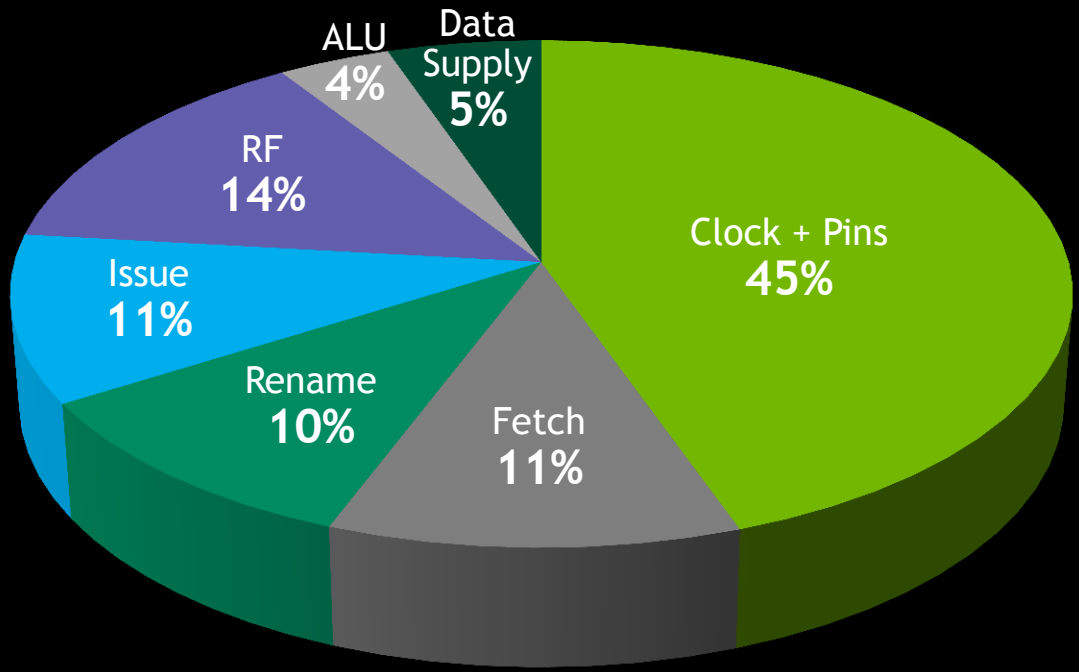
How is Power Spent in a CPU?

In-order Embedded

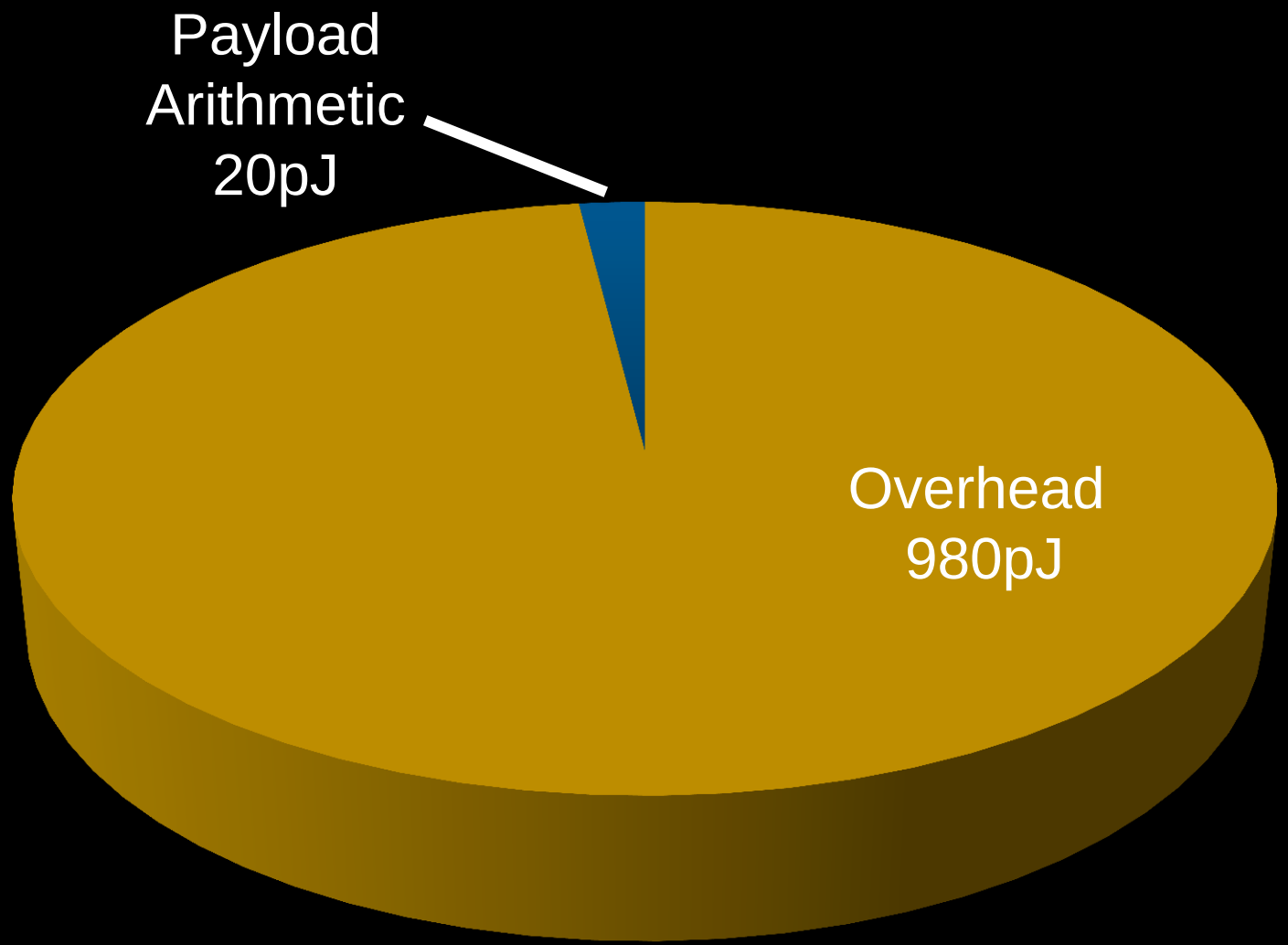


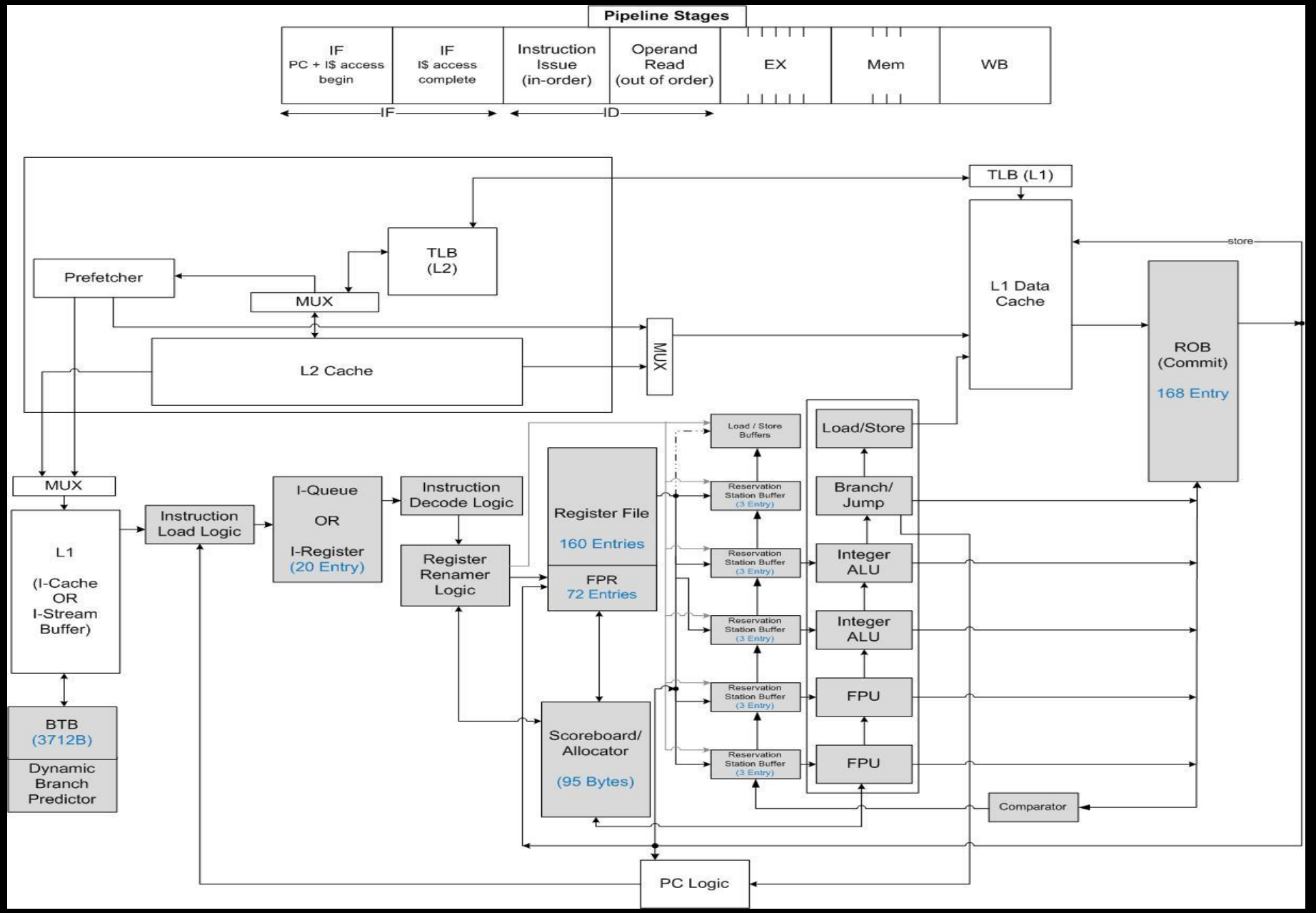
Dally [2008] (Embedded in-order CPU)

OOO Hi-perf

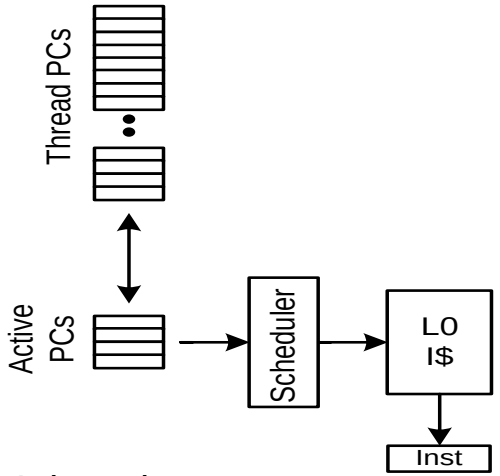


Natarajan [2003] (Alpha 21264)


Payload
Arithmetic
20pJ
Overhead
980pJ

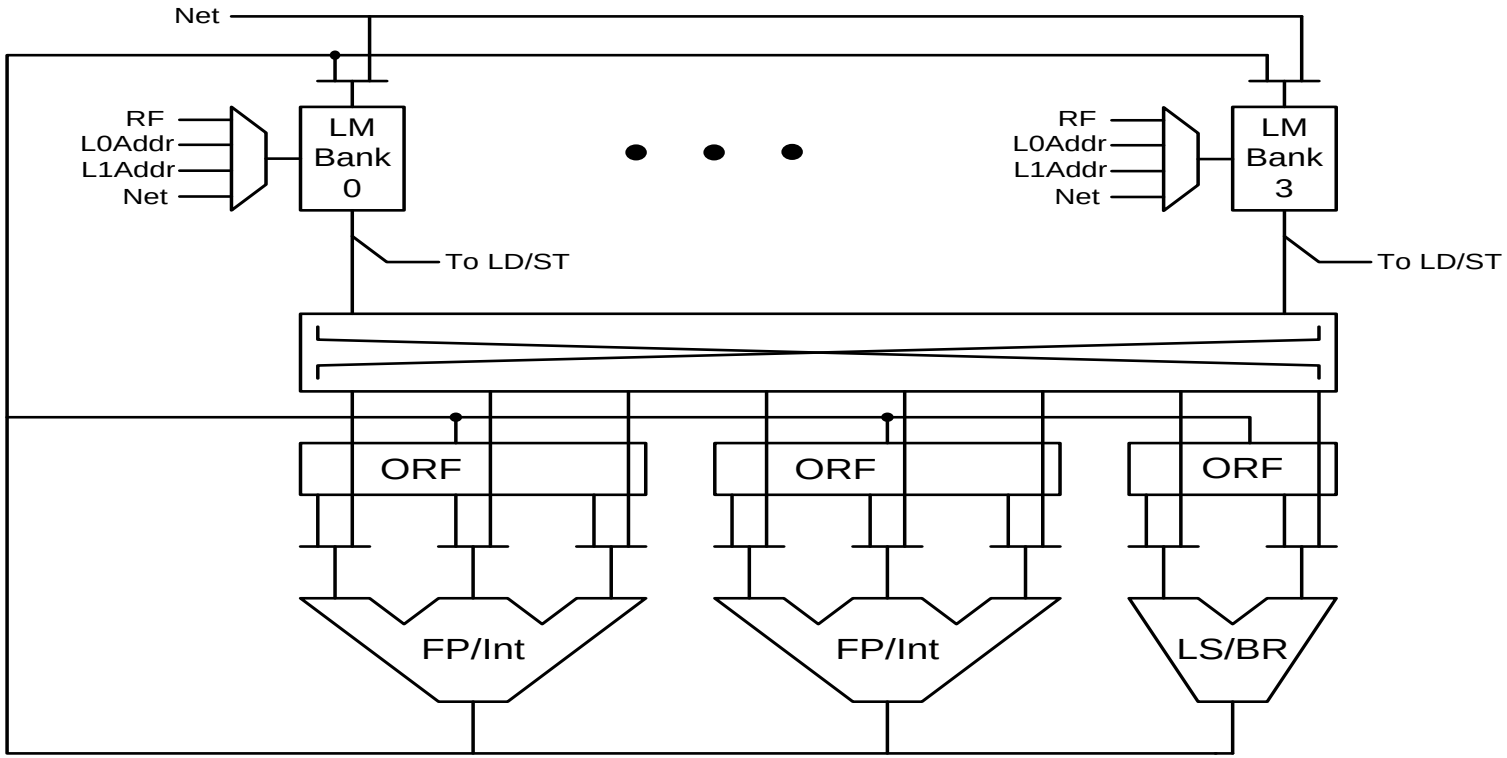

Pipeline Stages
IF
PC + I$ access begin
IF
I$ access complete
Instruction Issue (in-order)
Operand Read (out of order)
EX
Mem
WB
IF
ID
TLB (L1)
TLB (L2)
Prefetcher
MUX
L2 Cache
MUX
L1 Data Cache
store
ROB (Commit)
168 Entry
MUX
L1 (I-Cache OR I-Stream Buffer)
Instruction Load Logic
I-Queue OR I-Register (20 Entry)
Instruction Decode Logic
Register Renamer Logic
Register File
160 Entries
FPR
72 Entries
Load / Store Buffers
Reservation Station Buffer (3 Entry)
Load/Store
Branch/ Jump
Integer ALU
Integer ALU
FPU
FPU
BTB (3712B)
Dynamic Branch Predictor
Scoreboard/ Allocator
(95 Bytes)
Comparator
PC Logic

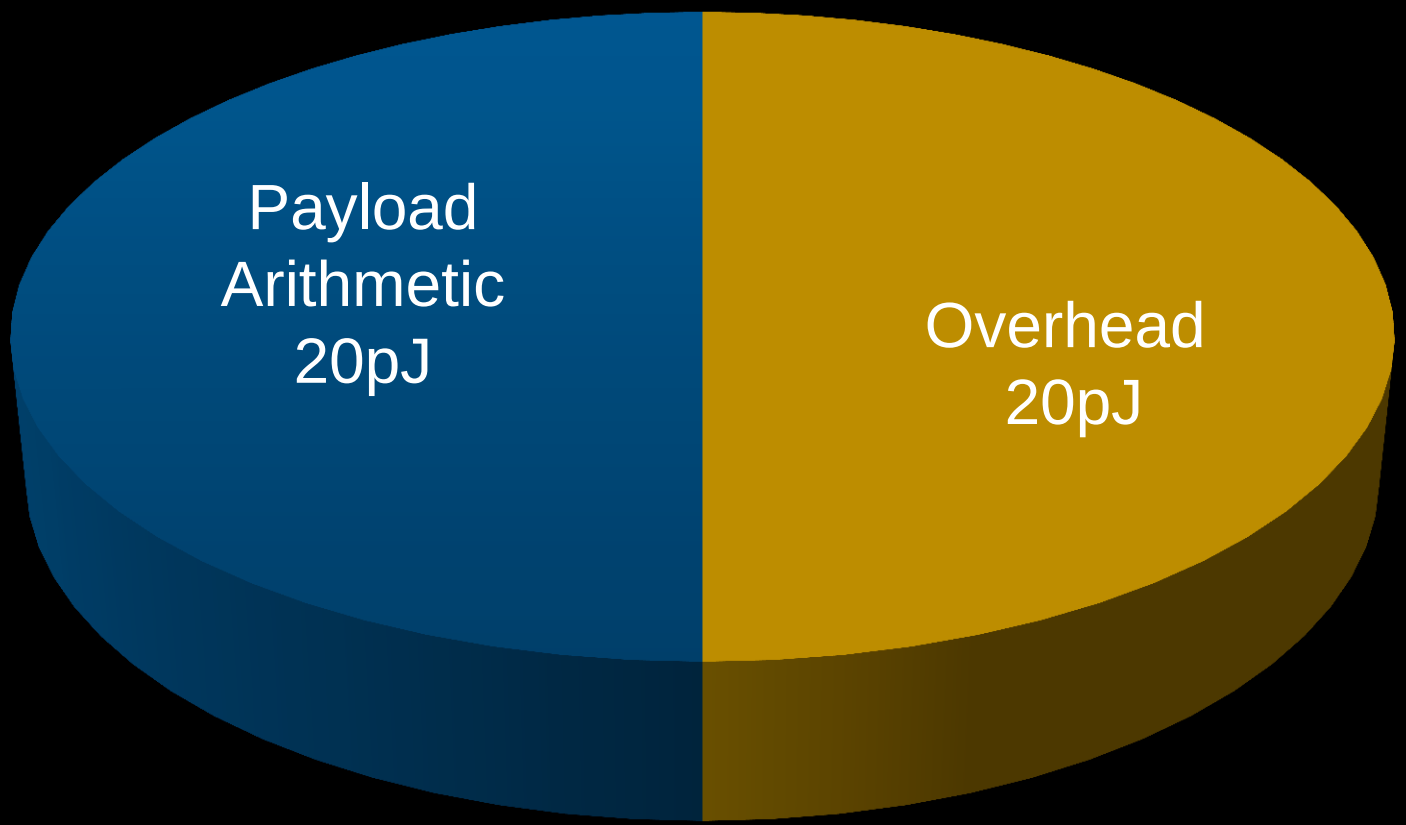
Control Path



64 threads
4 active threads
2 DFMAs (4 FLOPS/clock)
ORF bank: 16 entries (128 Bytes)
L0 I$: 64 instructions (1KByte)
LM Bank: 8KB (32KB total)

Data Path

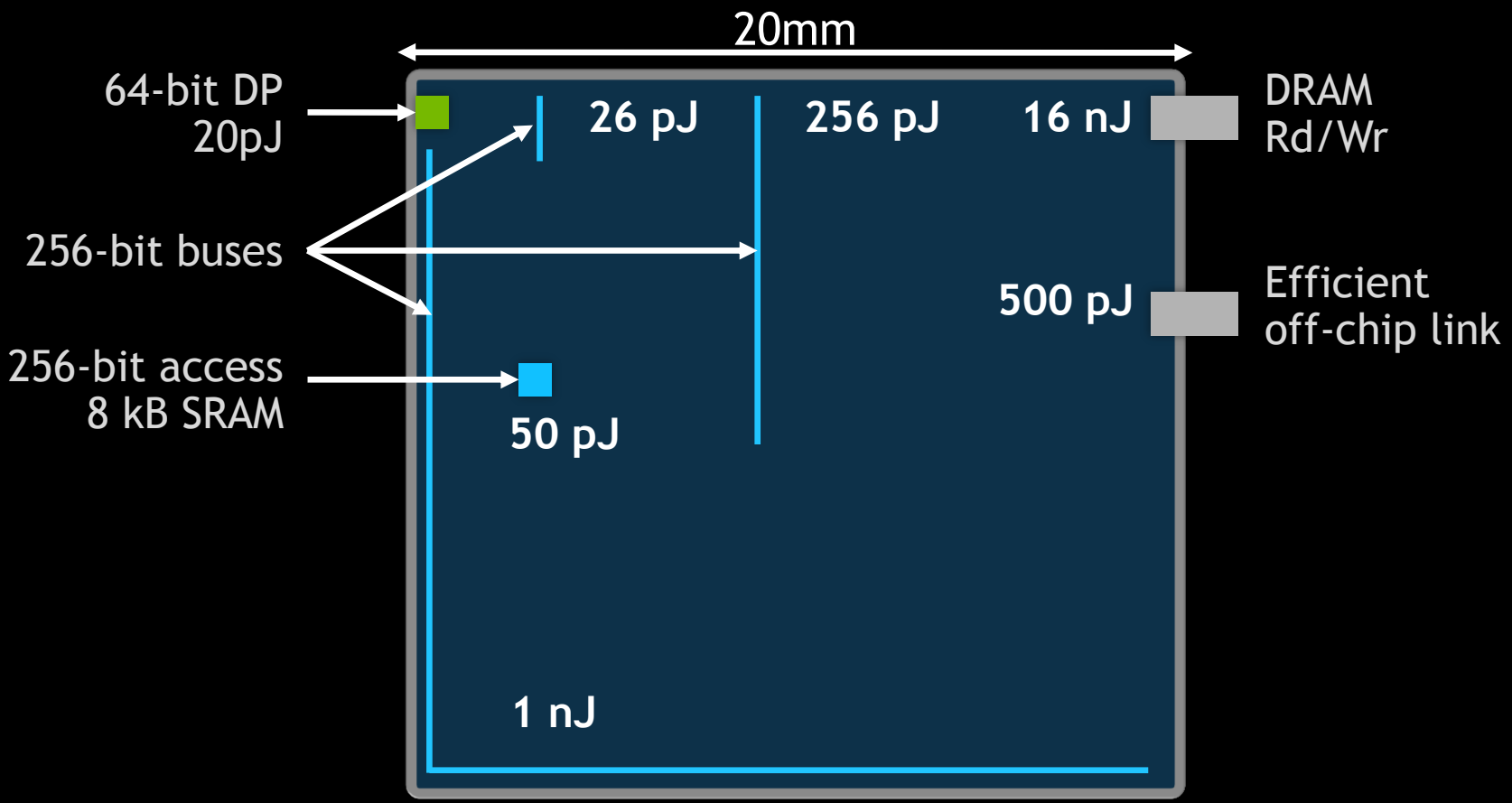



Payload
Arithmetic
20pJ
Overhead
20pJ

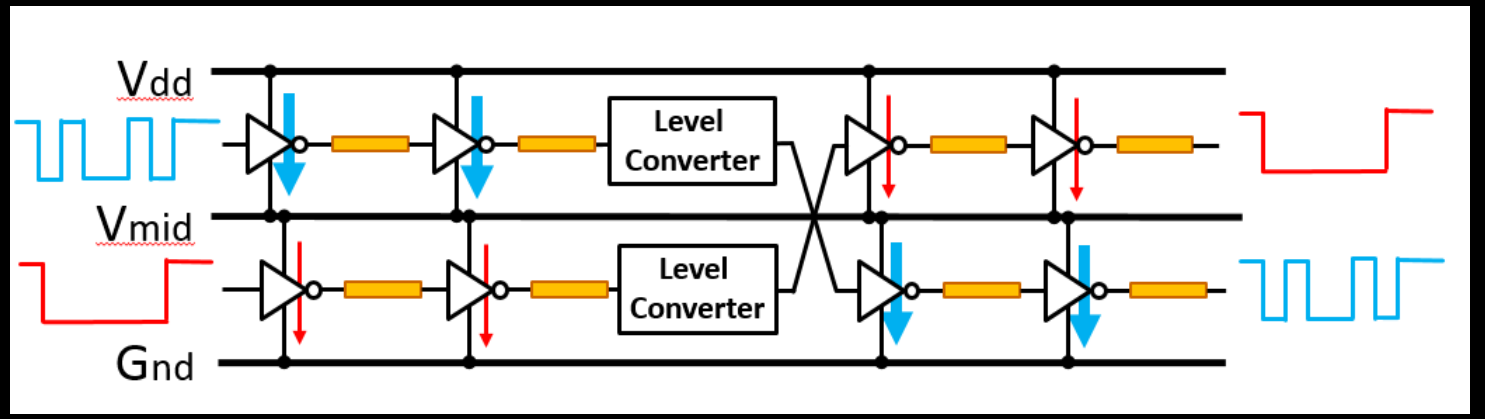
Energy-Efficient Architecture

- See Steve Keckler's Booth Talk – Wednesday 2:30PM
- How to reduce energy 10x when process gives 2x
  - Do Less Work
  - Eliminate redundancy, waste, and overhead
  - Move fewer bits – over less distance
  - Move data more efficiently

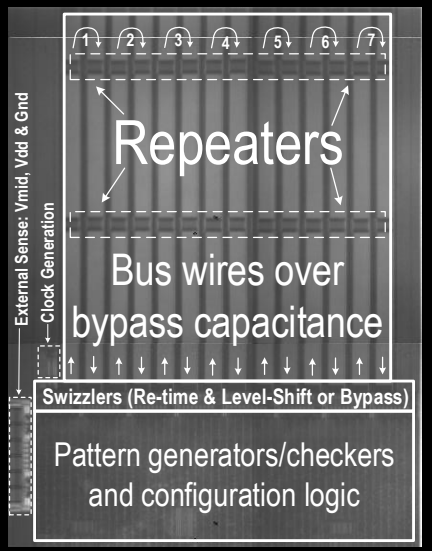
Communication Energy

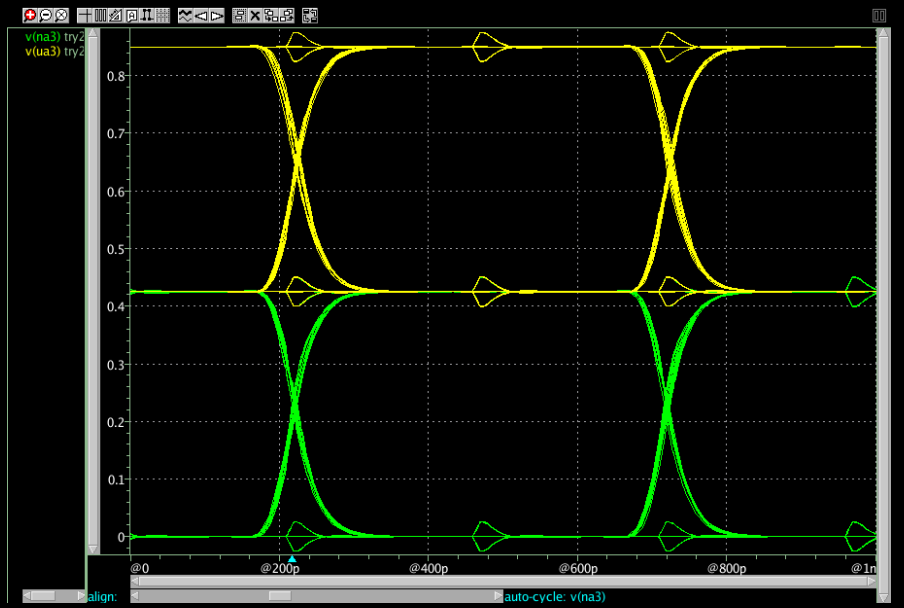


Charge-Recycled Signaling (CRS)

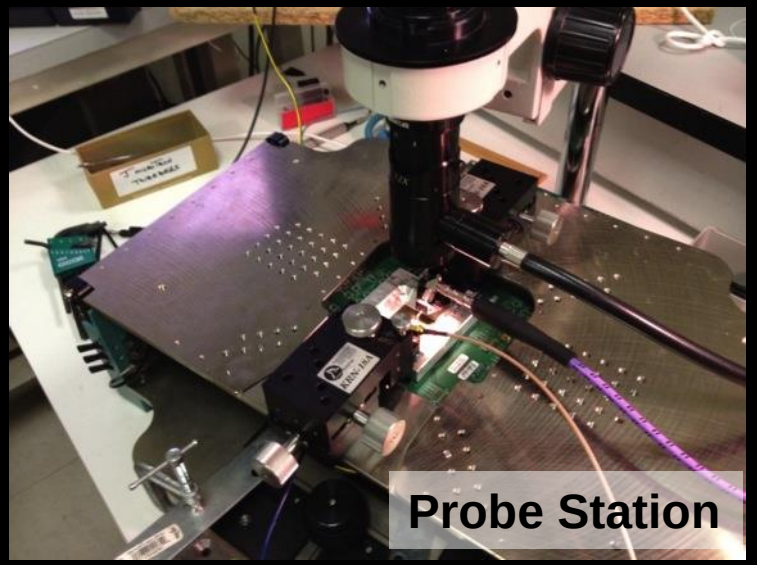


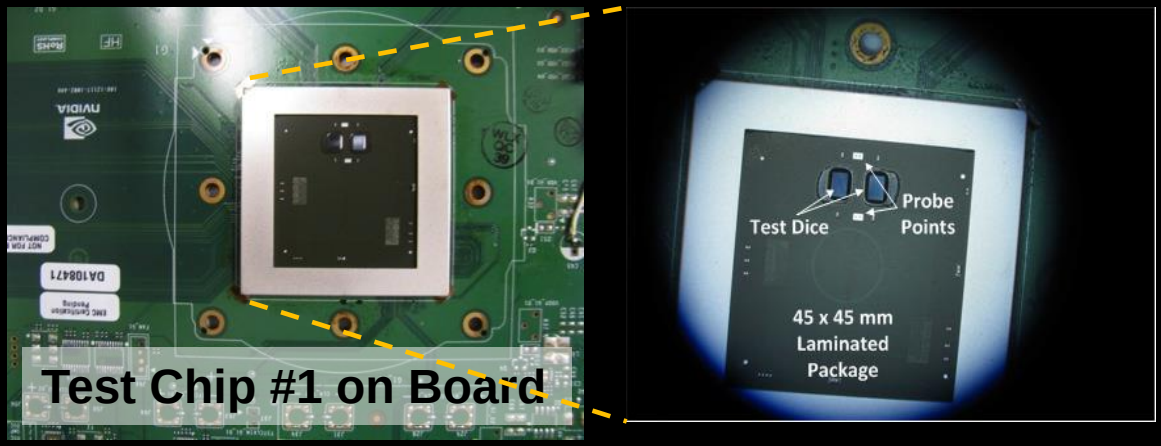
Reduces on-chip signaling energy by 4x

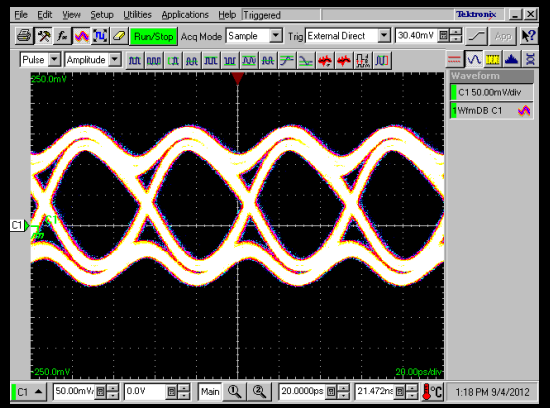




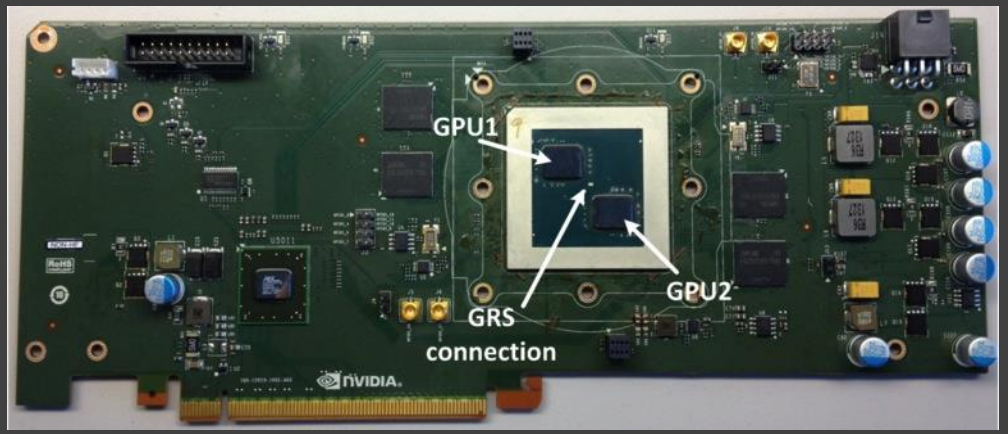
Ground-Referenced Signaling (GRS)







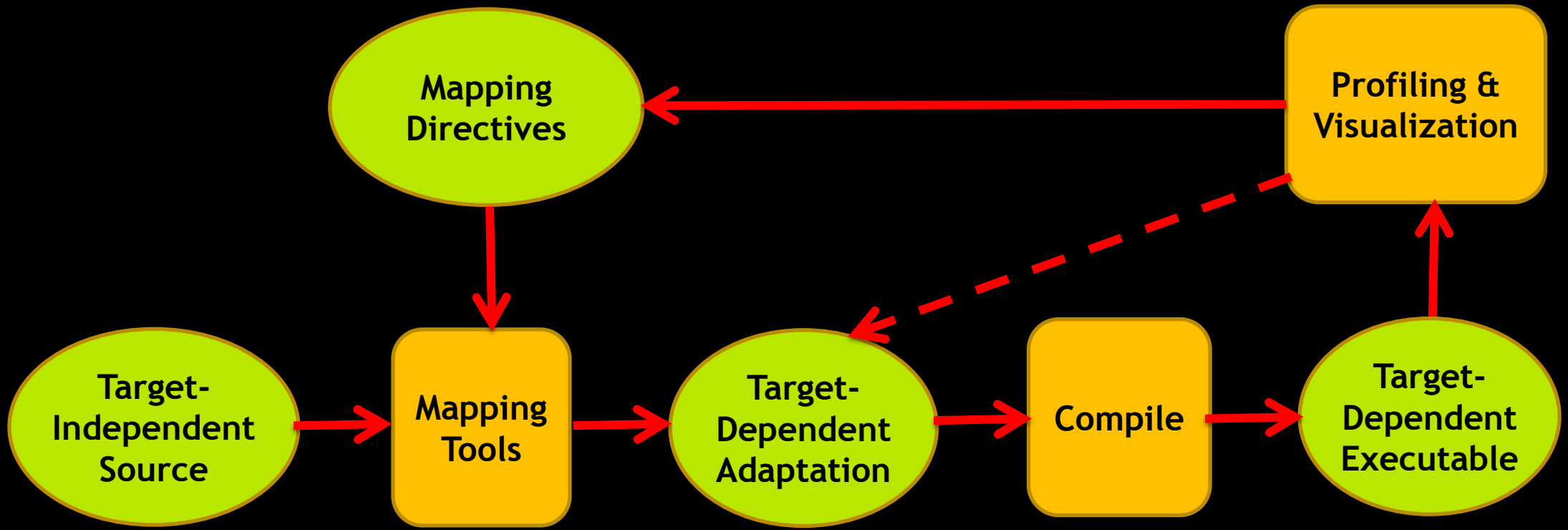
Eye Diagram from Probe



Poulton et al. ISSCC 2013, JSSCC Dec 2013

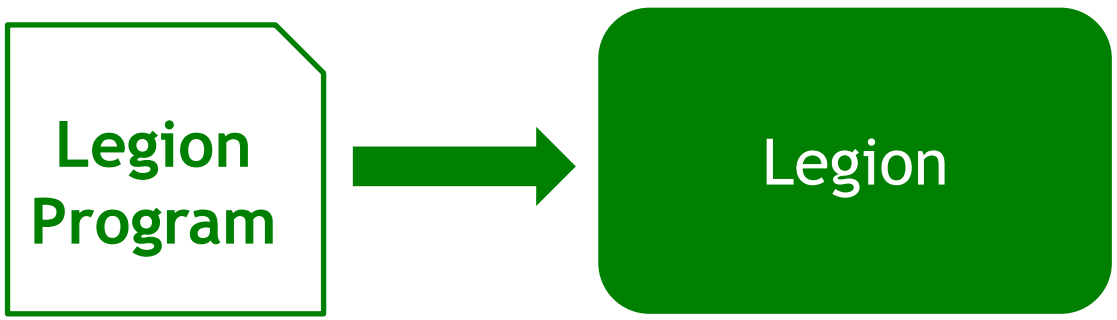
Programmability

Target-Independent Programming



Legion Programming Model

Enabling Powerful Program Analysis

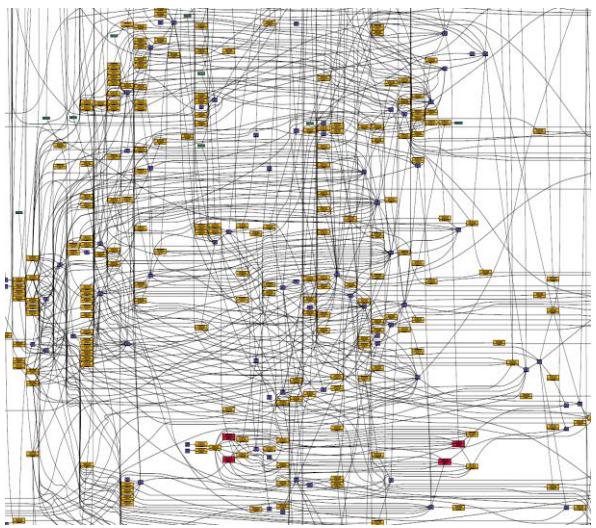


Machine-Independent Specification

Tasks: decouple control from machine

Logical regions: decouple program data from machine

Sequential semantics



Why it matters

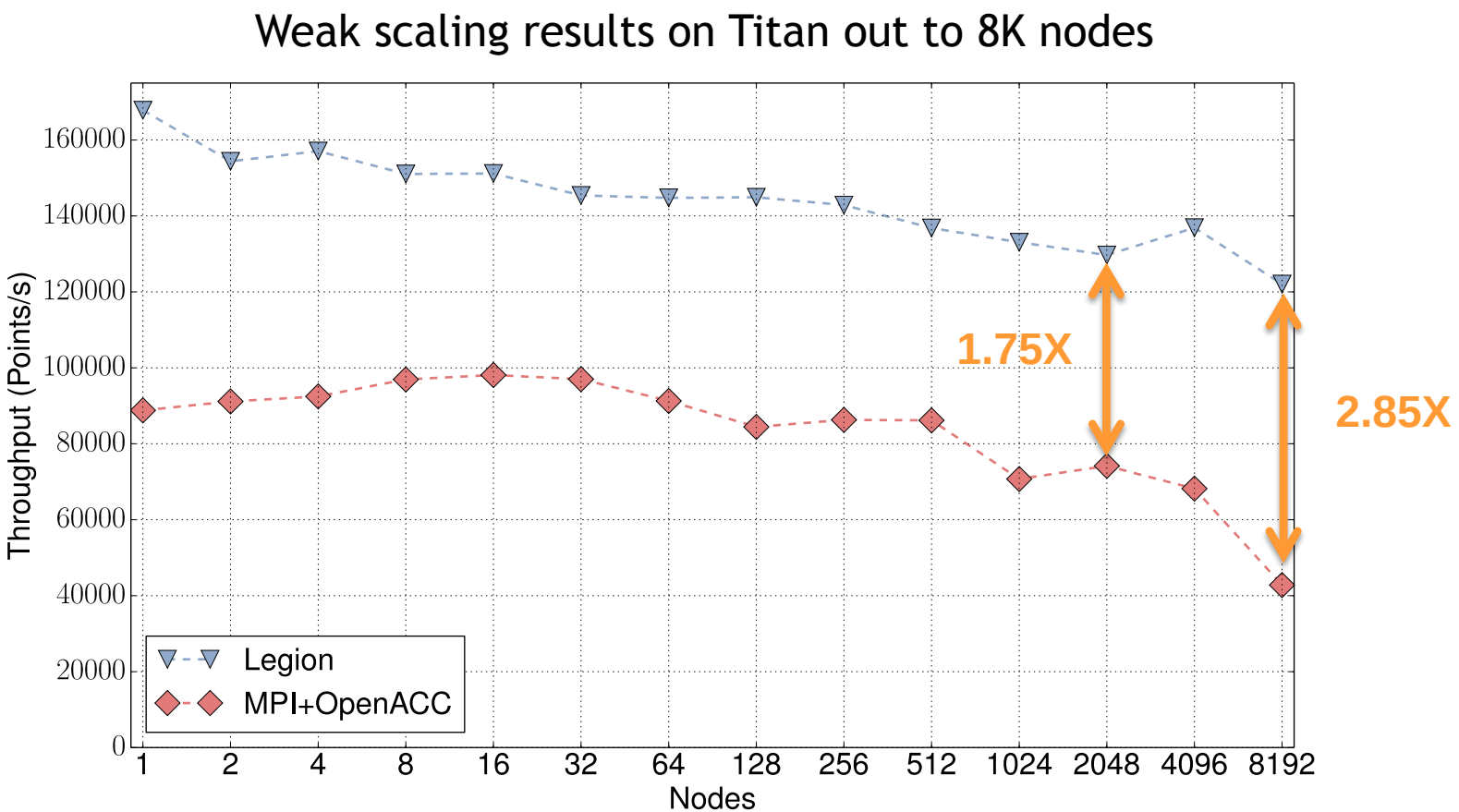
Reduce programmer pain

Extract ALL parallelism

Easily transform and remap programs for new machines

Comparison with MPI+OpenACC

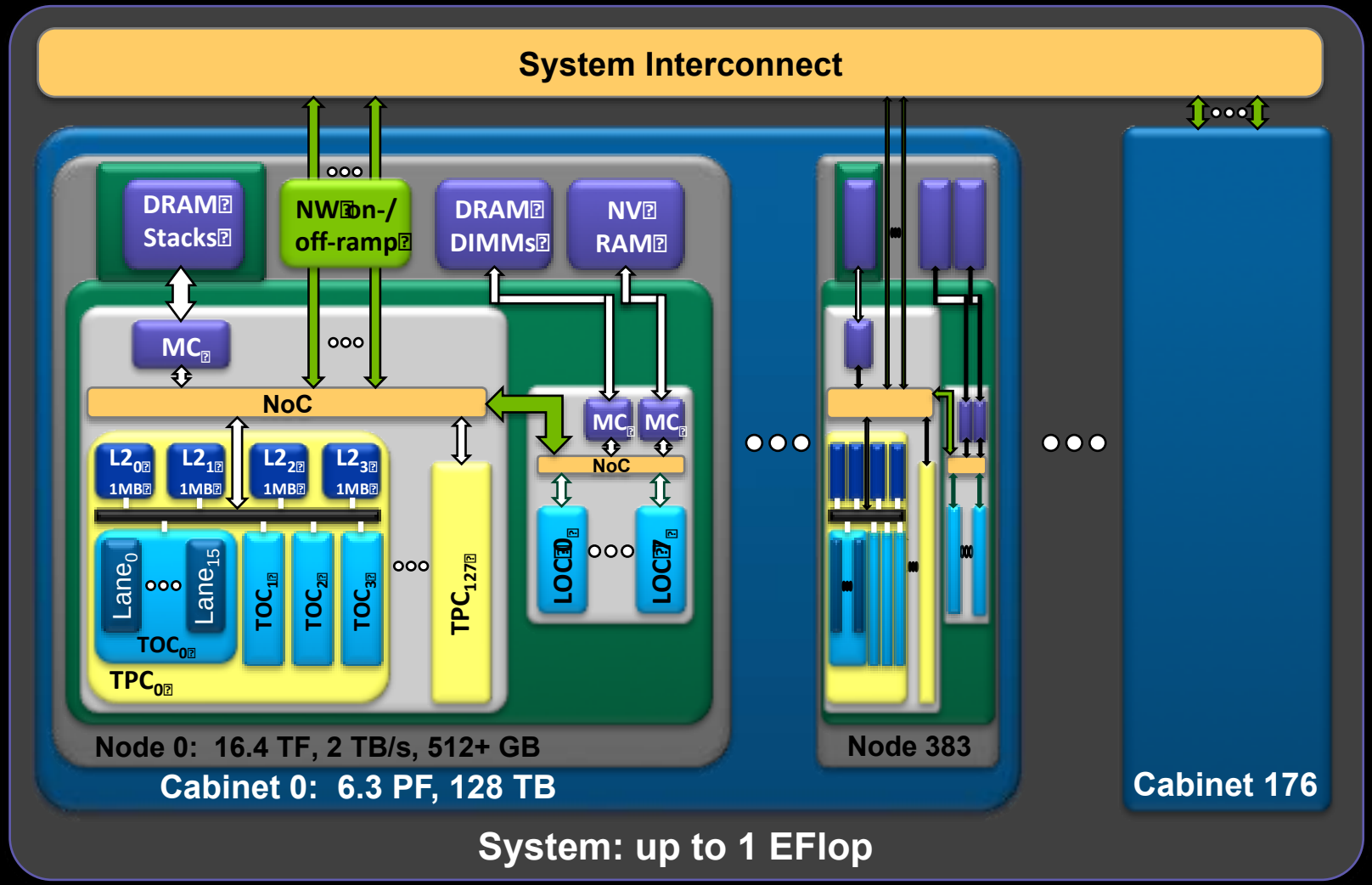
The power of program analysis



As application and machine complexity increases, the performance gap will grow.

Scalability

System Sketch

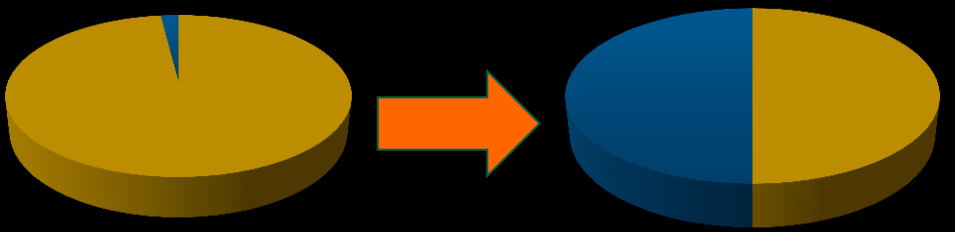


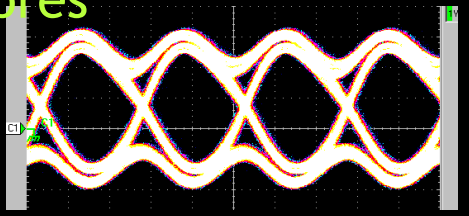
Heterogenous Network Requirements

- GPUs present unique requirements on network
  - 10^4 – 10^5 threads initiating transactions
  - Can saturate 150GB/s NVLINK bandwidth
- In addition to HPC requirements not met by commodity networks
  - Scalable BW up to 200GB/s per endpoint
  - <1us end-to-end latency at 16K endpoints
  - Scale to 128K endpoints
  - Load balanced routing
  - Congestion control
- Operations: Load/Store, Atomics, Messages, Collectives

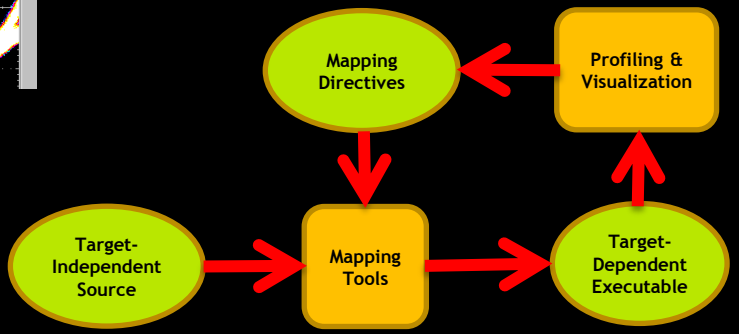
Conclusion

- Energy Efficiency
  - Reduce overhead with Throughput cores
  - Efficient Signaling Circuits
  - Enhanced Locality

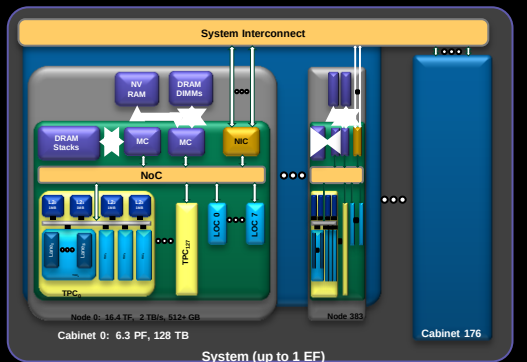


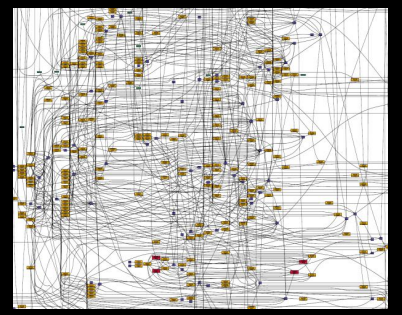


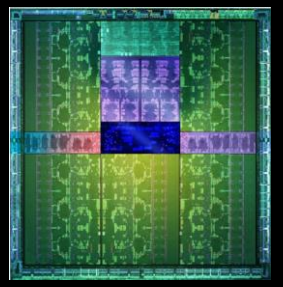
- Programming 10^{10} Threads
  - Target-independent programming – mapping via tools



- System Sketch
  - Efficient Nodes
  - GPU-Centric network







2013

20PF
18,000 GPUs
10MW
2 GFLOPs/W
~10^7 Threads

CORAL
150-300PF (5-10x)
11MW (1.1x)
14-27 GFLOPs/W (7-14x)
Lots of Threads

2017

You Are Here

2023

1,000PF (50x)
72,000HCNs (4x)
20MW (2x)
50 GFLOPs/W (25x)
~10^{10} Threads (1000x)