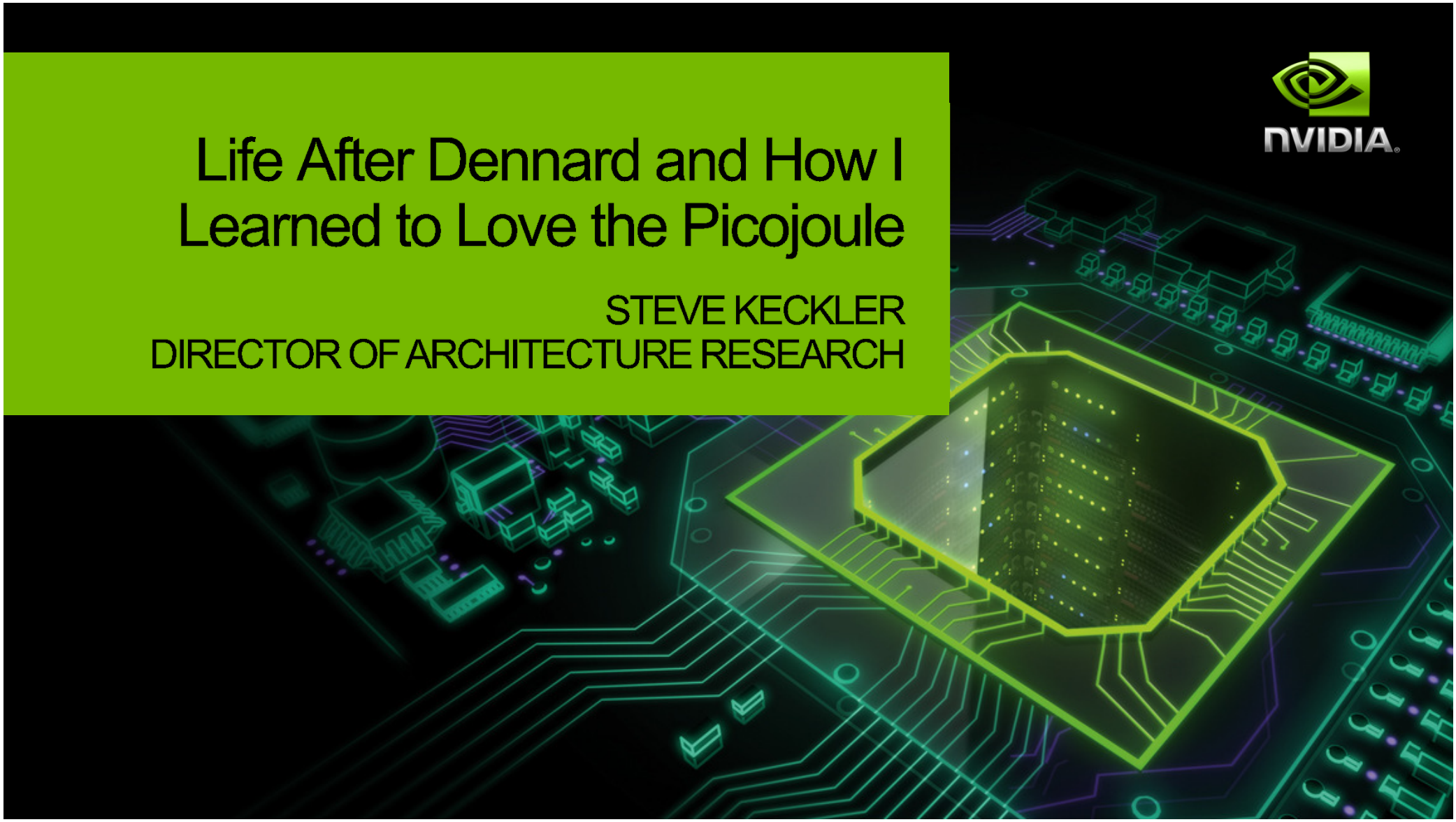
Life After Dennard and How I Learned to Love the Picojoule

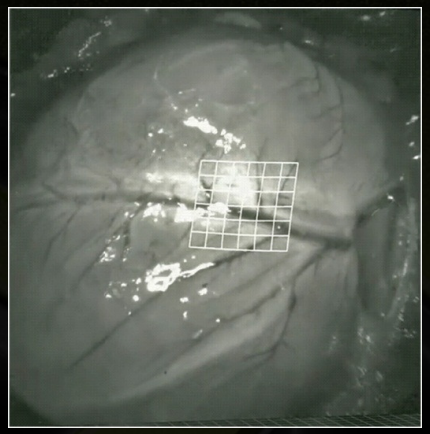
STEVE KECKLER
DIRECTOR OF ARCHITECTURE RESEARCH

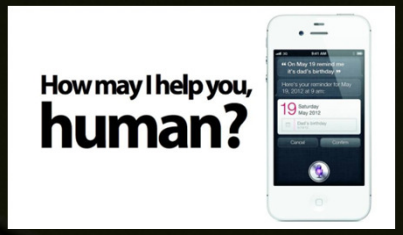


Performance-oriented Computing is Everywhere

© NVIDIA 2011

Providing Rich Immersive Experiences

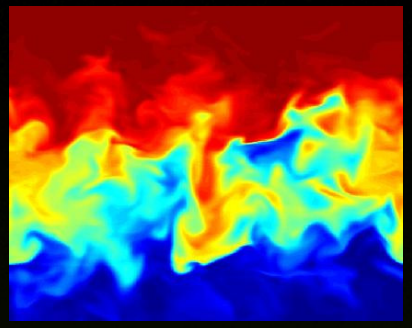




© NVIDIA 2011

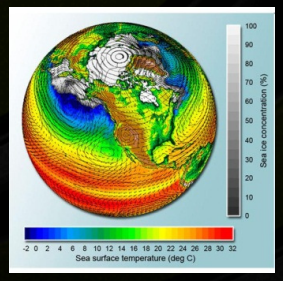
Addressing Societal Challenges

Product Design and Simulation

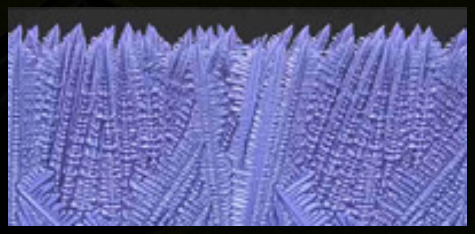


Alternative Fuels

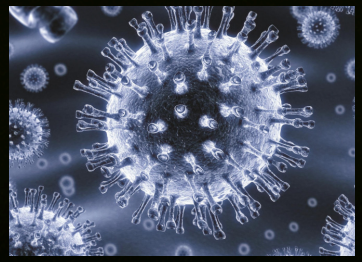
Medical Diagnosis



Climate Change



Materials Design



Understanding Viruses

© NVIDIA 2011

Insatiable Demand for Performance

Battlefield 1942
(2002)

Battlefield 3
(2011)

> 1 TFlop
~1KW

1W

© NVIDIA 2011

Which of These are Power Constrained?

© NVIDIA 2011

Silicon Scaling – the end of the easy ride

Where Does Processor Energy Go?

Strategies for Energy Efficiency

Challenges

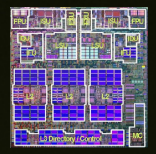
© NVIDIA 2011

Moore's Law is Only Part of the Story

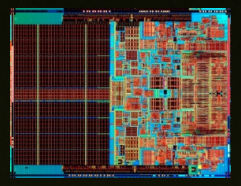
1993: 3M Tx

1997: 7.5M Tx

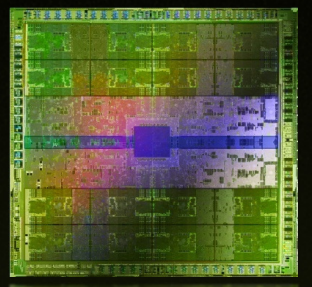
2001: 42M Tx



2004: 275M Tx



2007: 580M Tx

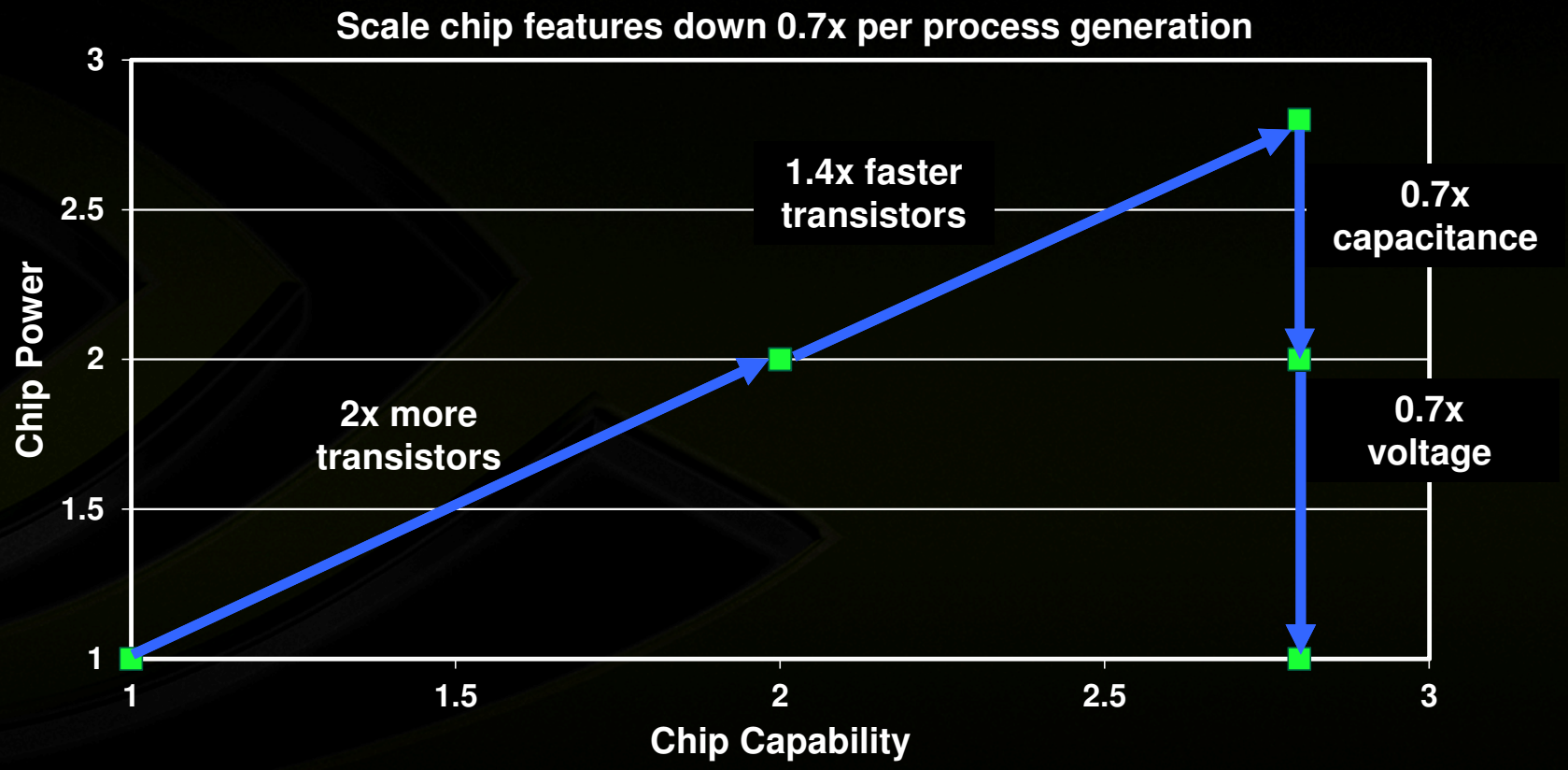


2010: 3B Tx

© NVIDIA 2011

Classic Dennard Scaling

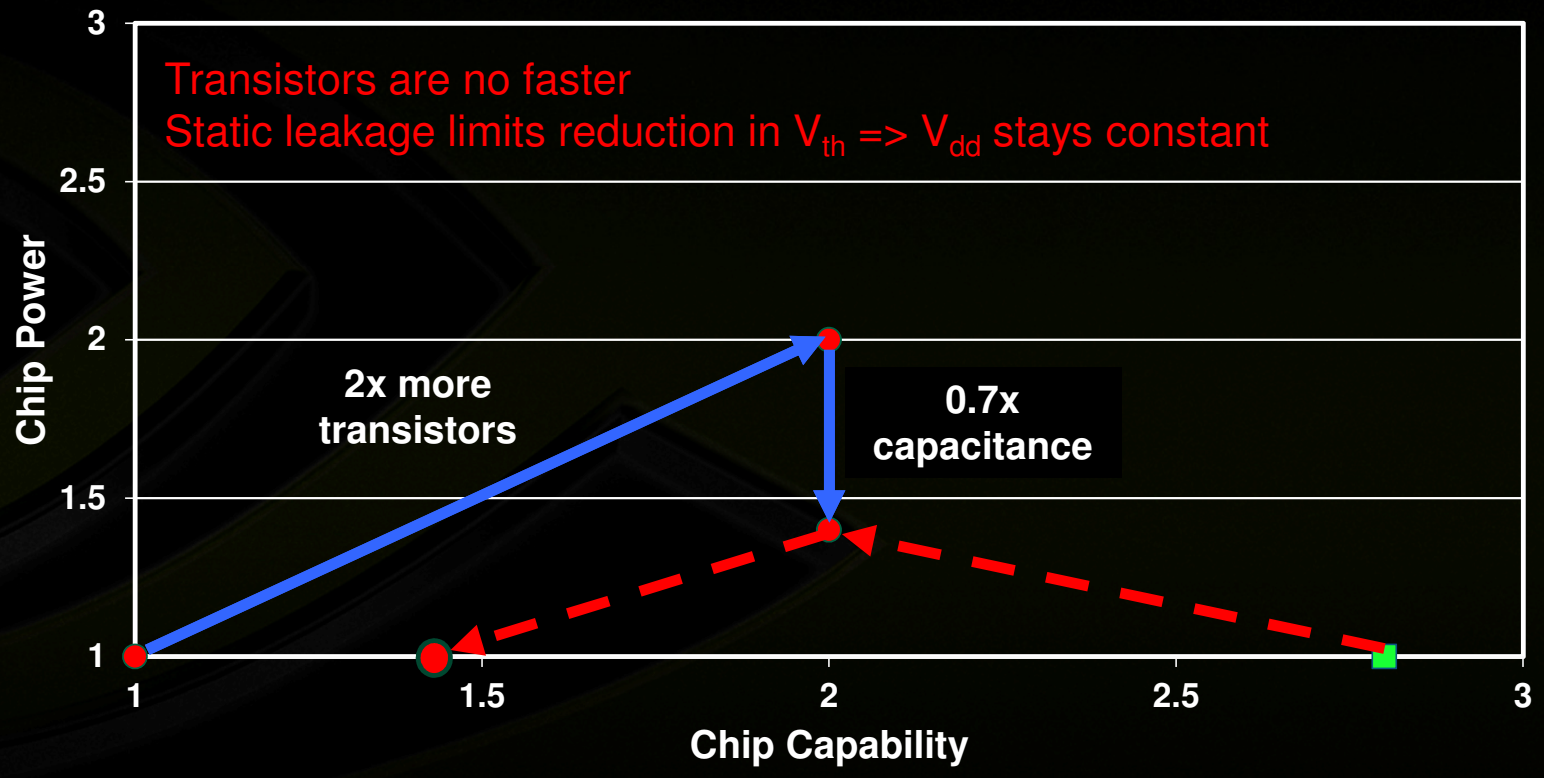
2.8x chip capability in same power



© NVIDIA 2011

Post Dennard Scaling

2x chip capability at 1.4x power
1.4x chip capability at same power



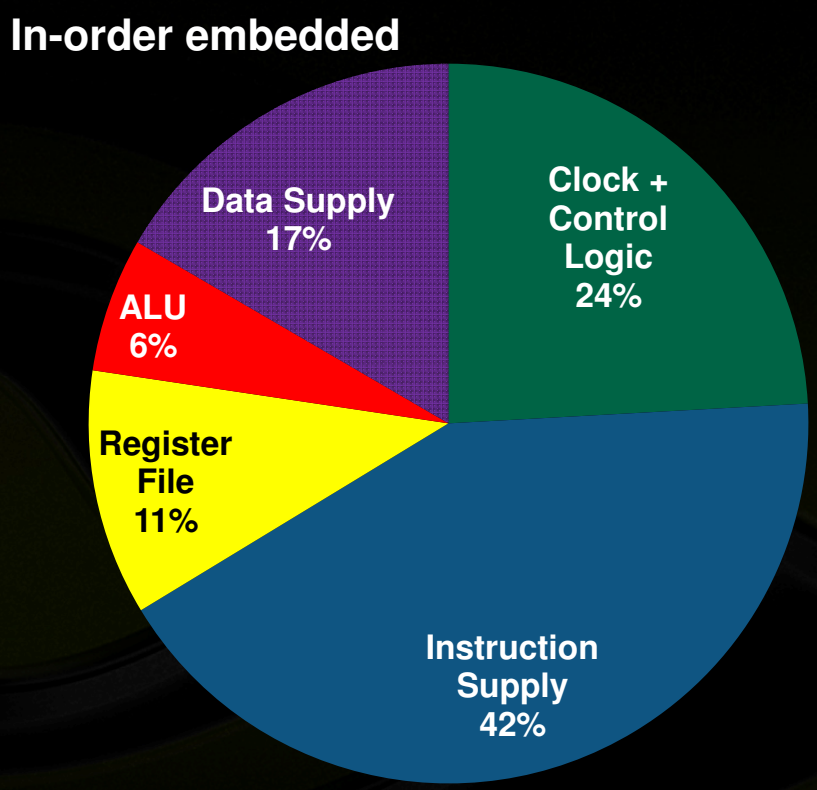
© NVIDIA 2011

Implications

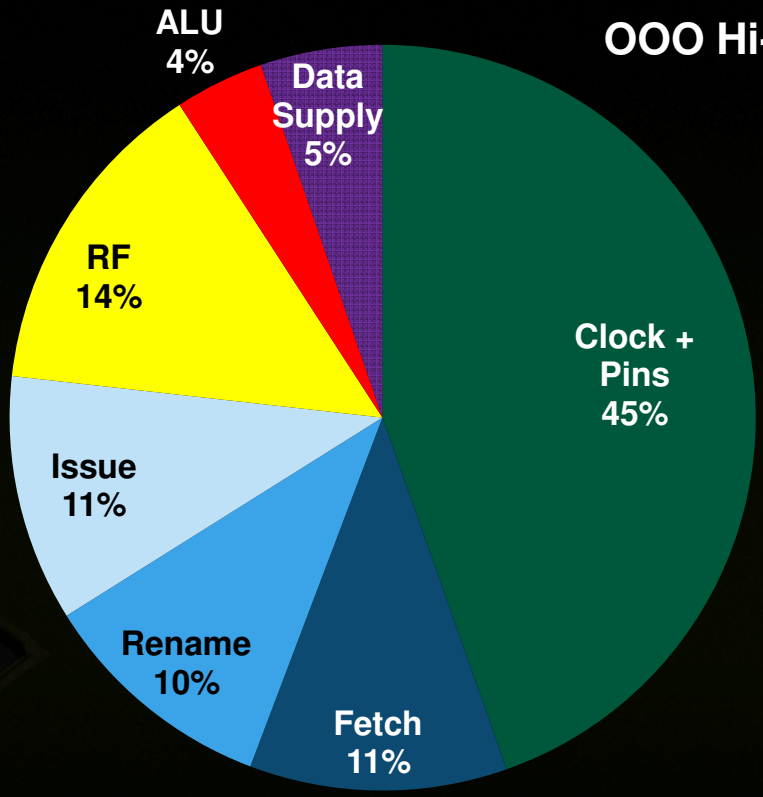
- End of clock-rate scaling => need more parallelism
- Capability drops down to 1.4 per generation.
  - Optimistic because wires scale worse than transistors
  - 32x gap in a decade from what continued Dennard scaling would give
- Scaling up conventional core count is not feasible
  - Best case – 40% more cores per generation
  - Far cry from 2x performance per generation
- Need better materials, devices, circuits, architectures, and SW

© NVIDIA 2011

How is power spent in a CPU?



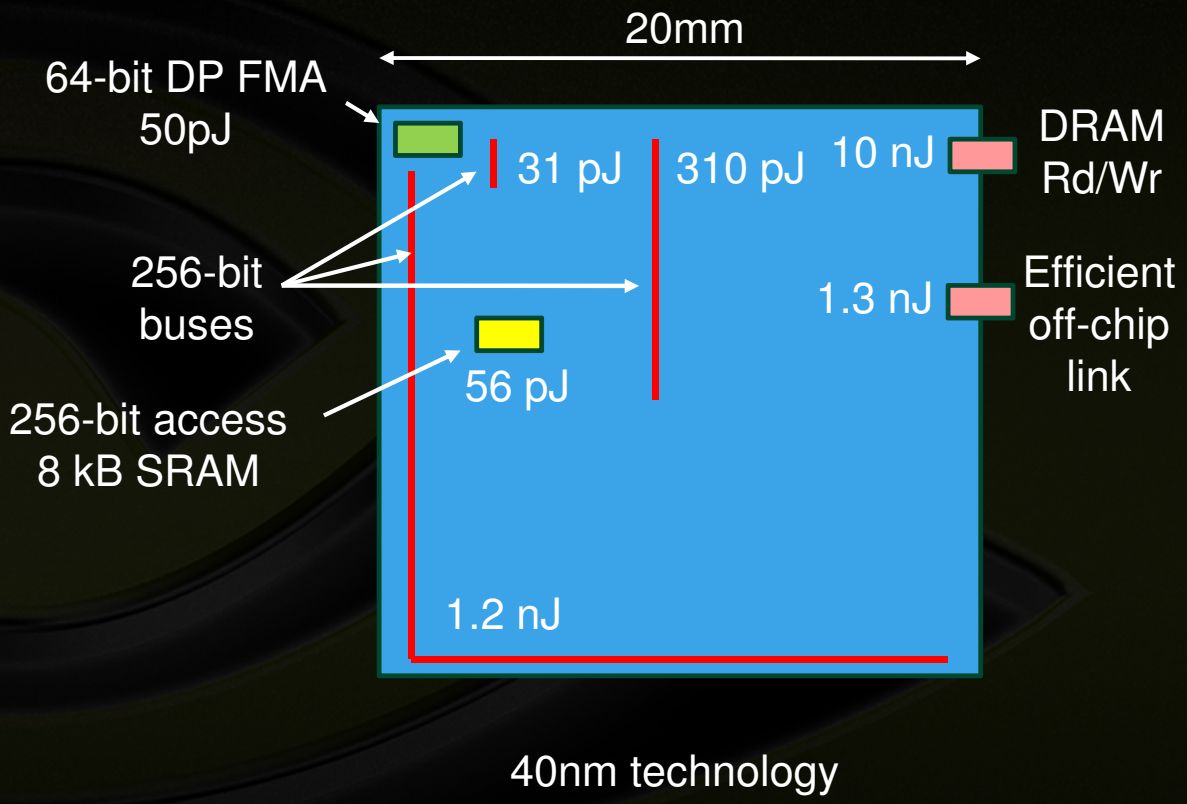
Dally [2008] (Embedded in-order CPU)



Natarajan [2003] (Alpha 21264)

© NVIDIA 2011

Fundamental Energy Costs



Today's high perf CPUs

- **1-2 nJ/instruction**

20-40x energy overhead

- **Instruction control**
- **Data movement**

© NVIDIA 2011

Energy Shopping List

Processor Technology	40 nm	10nm
Vdd (nominal)	0.9 V	0.7 V
DFMA energy	50 pJ	7.6 pJ
64b 8 KB SRAM Rd	14 pJ	2.1 pJ
Wire energy (256 bits, 10mm)	310 pJ	174 pJ

Memory Technology	45 nm	16nm
DRAM interface pin bandwidth	4 Gbps	50 Gbps
DRAM interface energy	20-30 pJ/bit	2 pJ/bit
DRAM access energy	8-15 pJ/bit	2.5 pJ/bit

FP Op lower bound
=
8 pJ

Vogelsang [Micro 2010]

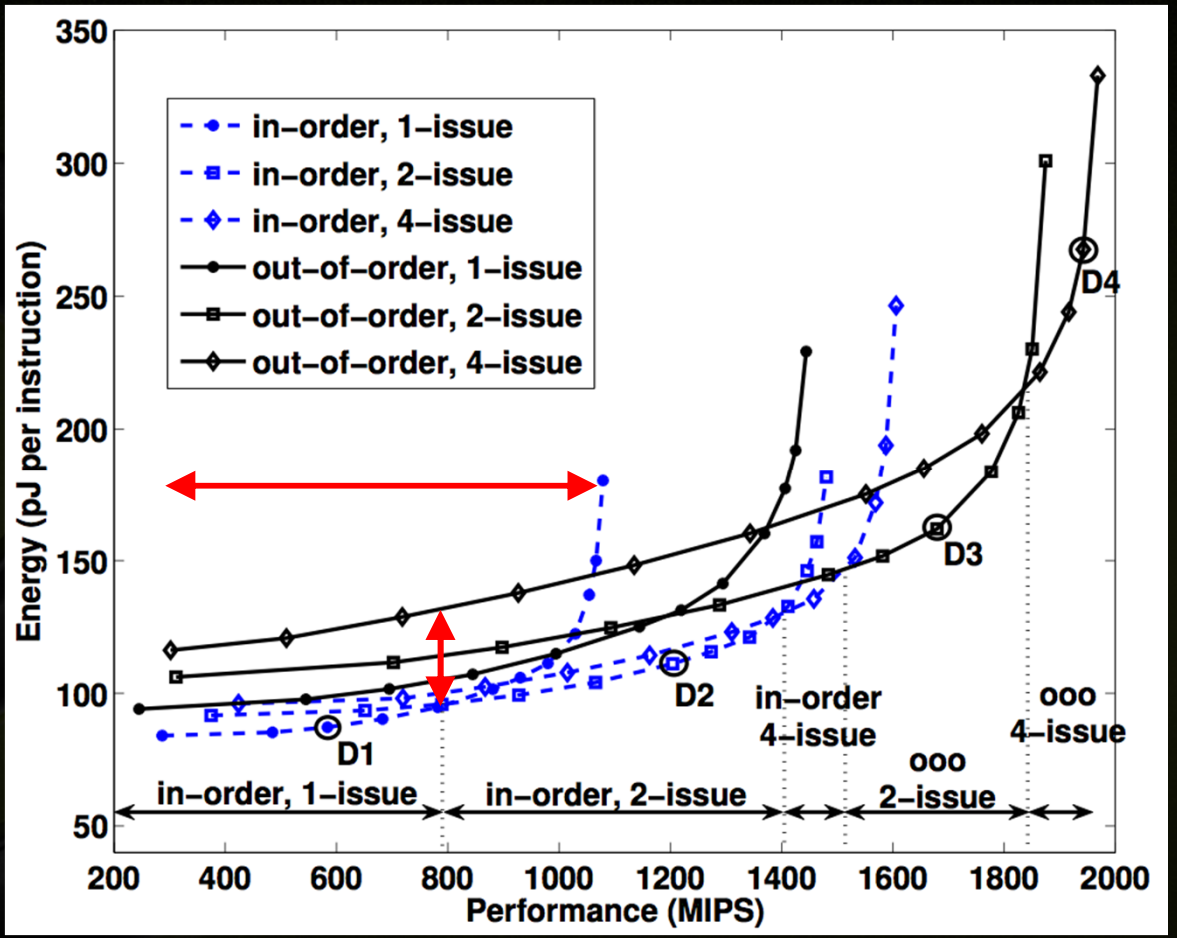
© NVIDIA 2011

Strategies for Energy Reduction

- Simpler processor architectures
- Reduce waste
- Improve physical locality of data
- Exploit heterogeneity and specialization
- Signaling and packaging
- Inch down voltage

© NVIDIA 2011

Simpler/Slower Cores = Energy Efficiency



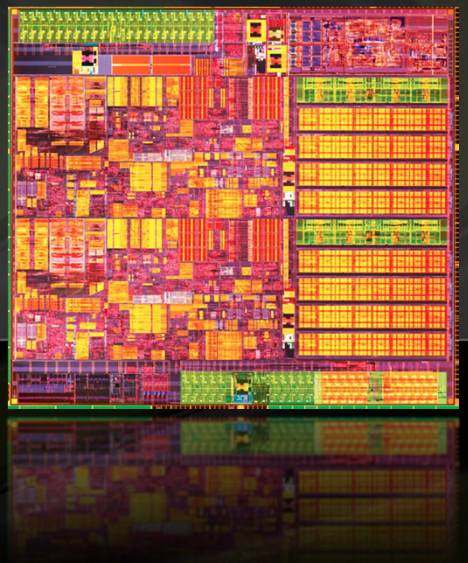
Azizi [PhD 2010]

© NVIDIA 2011

CPU
1.7 nJ/flop

Optimized for Latency

Caches

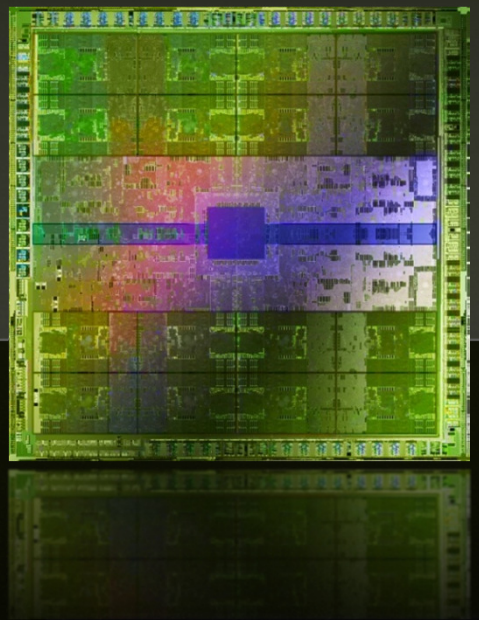


Westmere
32 nm

GPU
225 pJ/flop

Optimized for Throughput

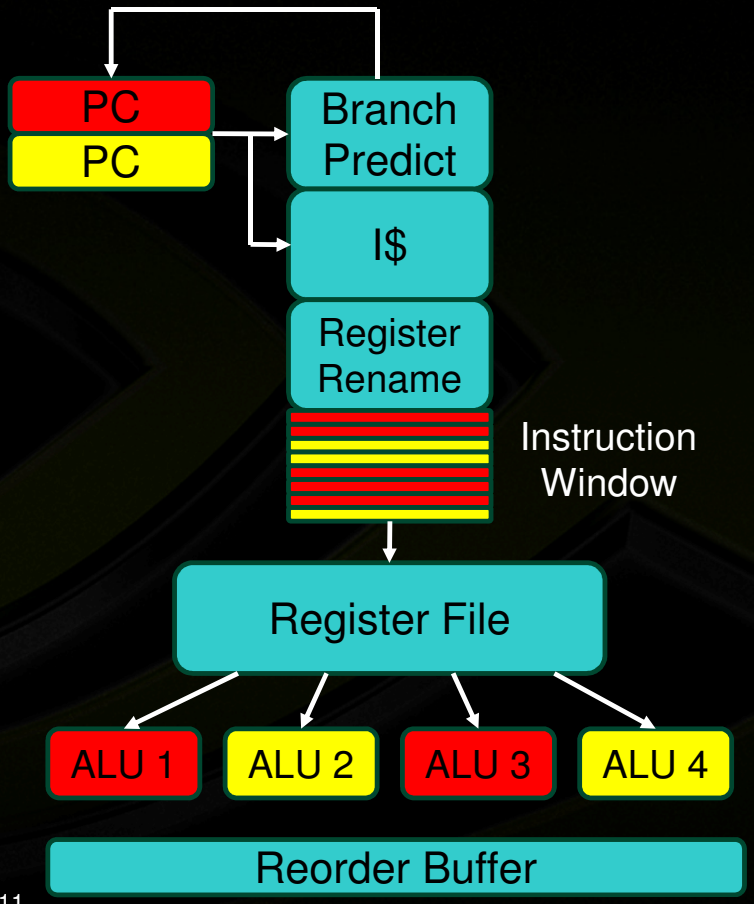
Explicit Management
of On-chip Memory



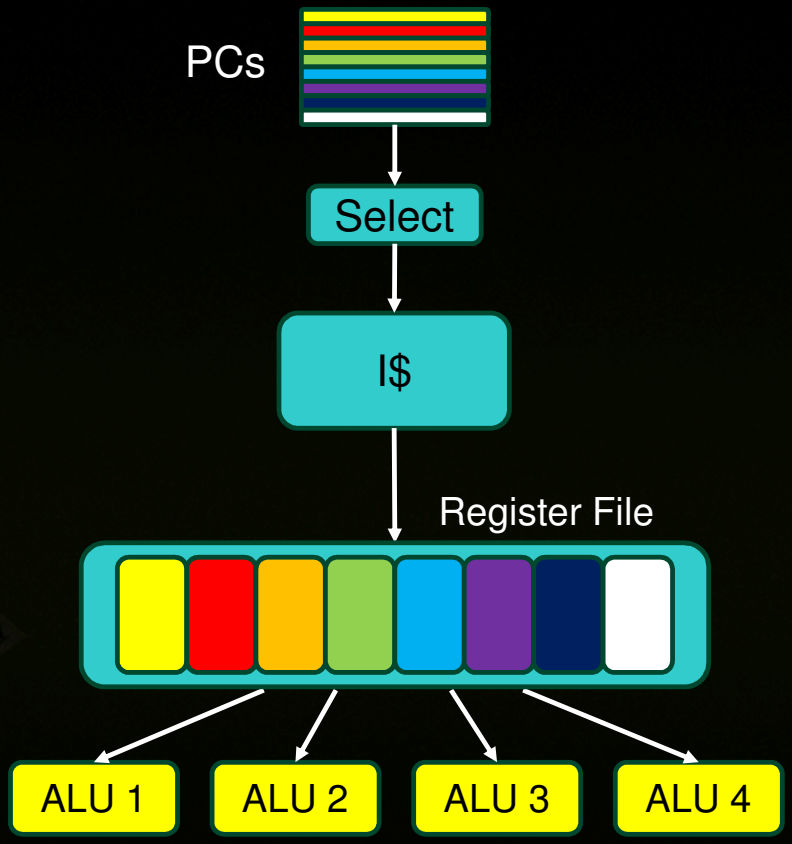
Fermi
40 nm

© NVIDIA 2011

Latency Processor



Throughput Processor

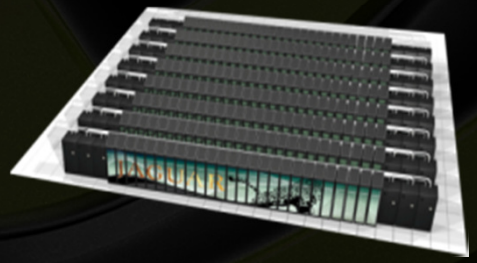


© NVIDIA 2011

Throughput Processors at Petascale

#1 : K Computer
68K Fujitsu Sparc CPUs
10.5 PFLOPS, 12.6MW

#2 : Tianhe-1A
7K Tesla GPUs
2.6 PFLOPS, 4MW



#3 : Jaguar
36K AMD Opteron CPUs
1.8 PFLOPS, 6.9MW

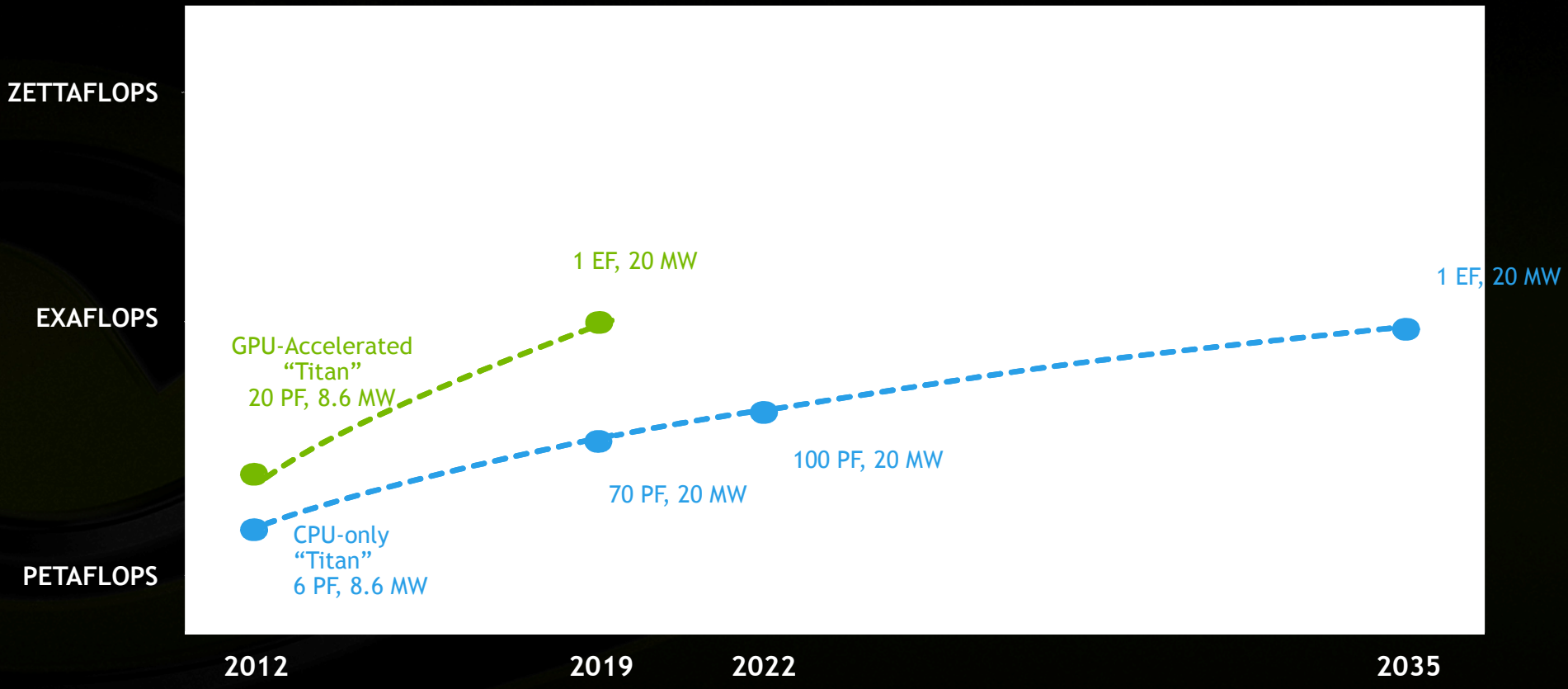
Titan
18K Tesla GPUs
>20 PFLOPS, 8.6MW

#4 : Nebulae
4K Tesla GPUs
1.3 PFLOPS, 2.5MW

#5 : Tsubame 2.0
4K Tesla GPUs
1.2 PFLOPS, 1.3MW
(most efficient PF system)

© NVIDIA 2011

The Road to Exascale



© NVIDIA 2011

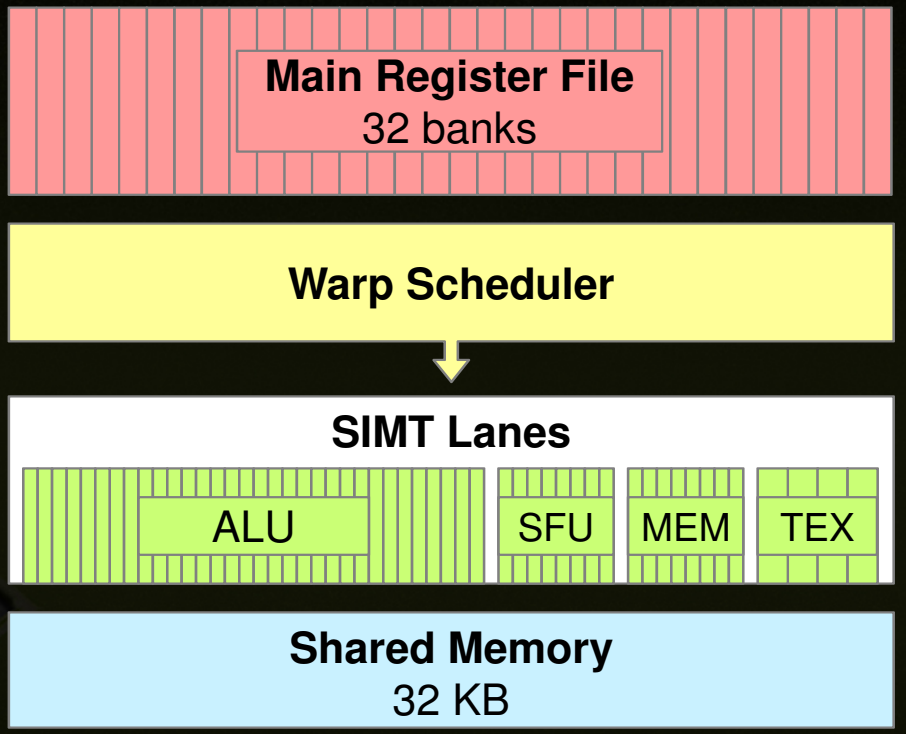
Throughput Processor Optimizations

- Hierarchical hardware thread scheduling
- Register File caching
- SIMT and Temporal SIMT execution

© NVIDIA 2011

Streaming Multiprocessor (SM)

- Multithreading with:
  - 32 threads per Warp
  - 48 Warps/SM
- Large register files
  - 1500+ threads/SM
  - Fermi: 128KB of RF
  - Heavily banked to provide high bandwidth
- Large operand read/write energy
  - ~15% of power of SM

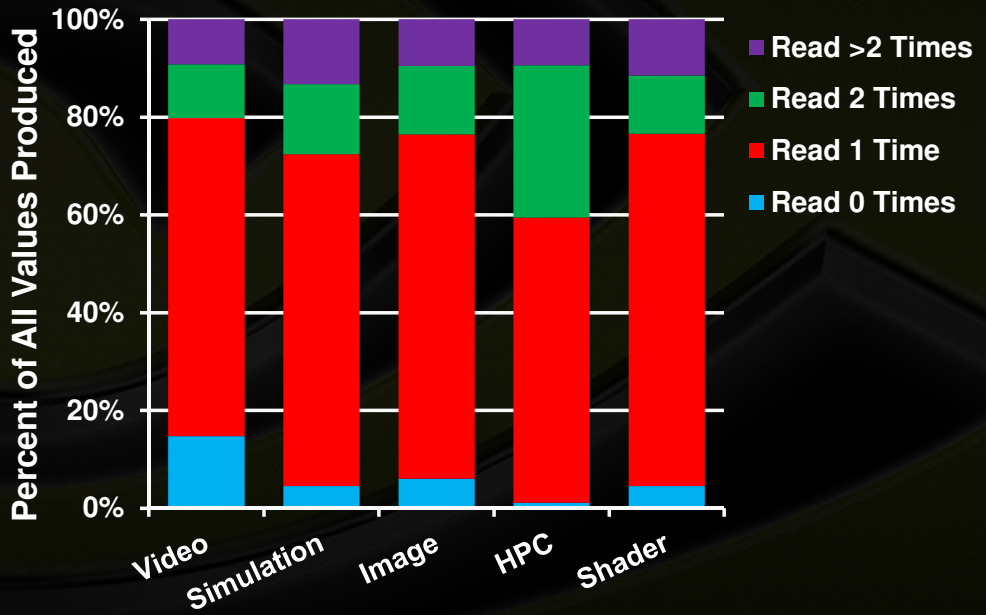


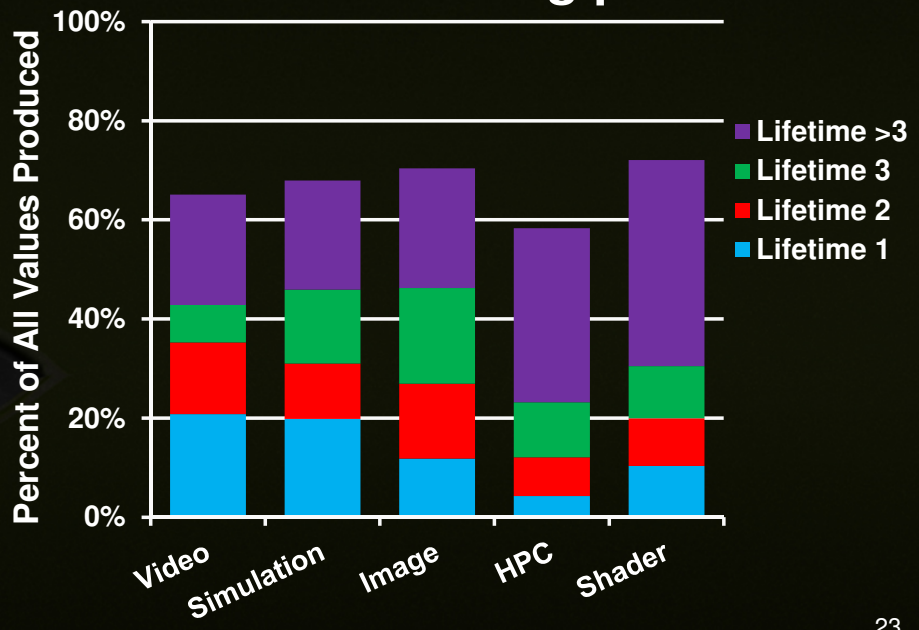
© NVIDIA 2011

Optimization Opportunities

- Large number of threads hide two types of latency
  - Long: global memory access (~400 cycles)
  - Short: ALU and shared memory access (8-20 cycles)

- 40% of values are read once, within 3 instructions of being produced

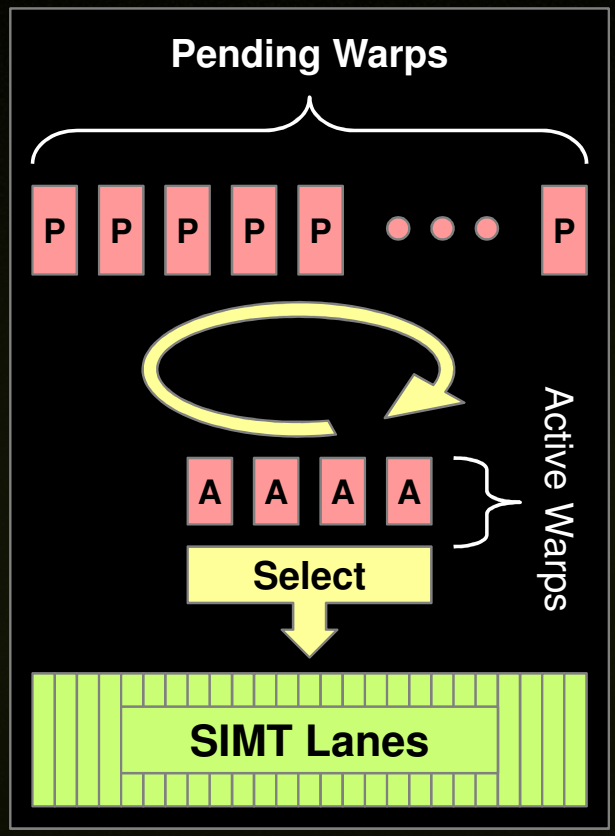




© NVIDIA 2011

Two-Level Scheduling

- Active warps used to tolerate short latency events
- Simplified instruction scheduler only considers active warps each cycle
- 8 active warps (of 48) is enough to sustain full SM throughput

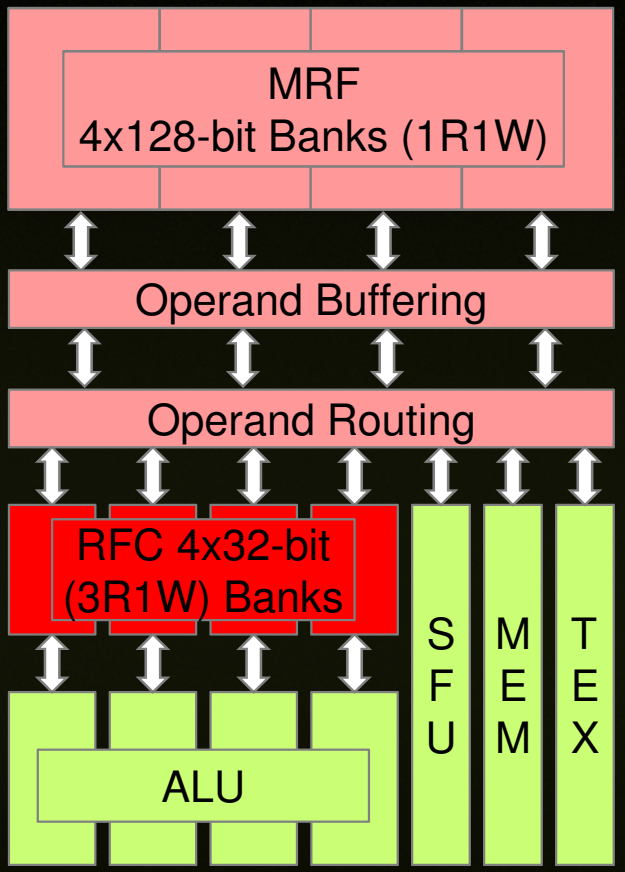


Gebhart, et. al [ISCA 2011, Micro 2011]

© NVIDIA 2011

Register File Caching (RFC)

- Private RFC per SIMT lane
  - 4-6 entries per thread
  - Operand routing when results are needed by shared units
- 48x smaller than MRF
  - When combined with 2-level scheduling
- Reduces up to 80% of RF reads/writes

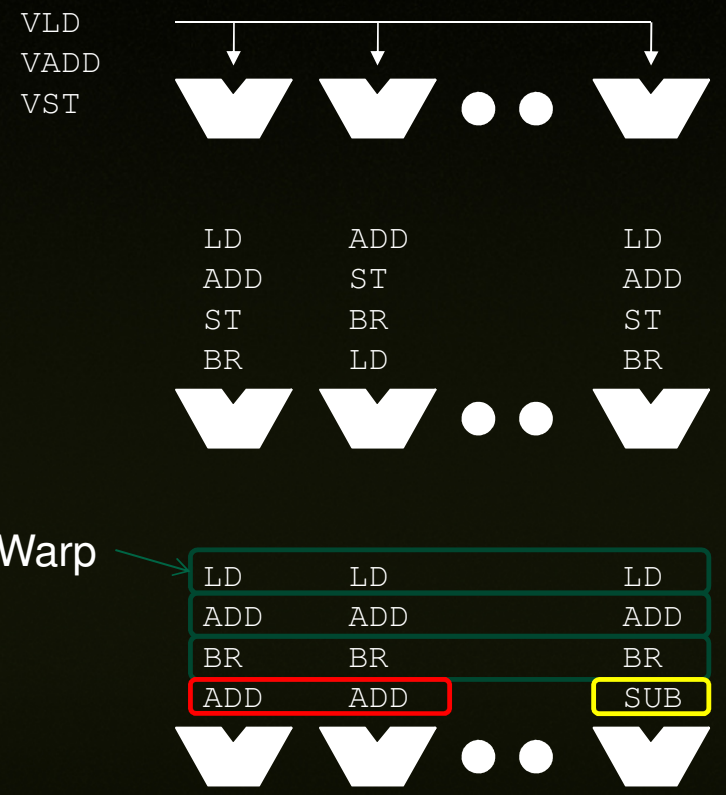


SM Cluster - replicated 8 times to form SM

© NVIDIA 2011

SIMD versus MIMD versus SIMT?

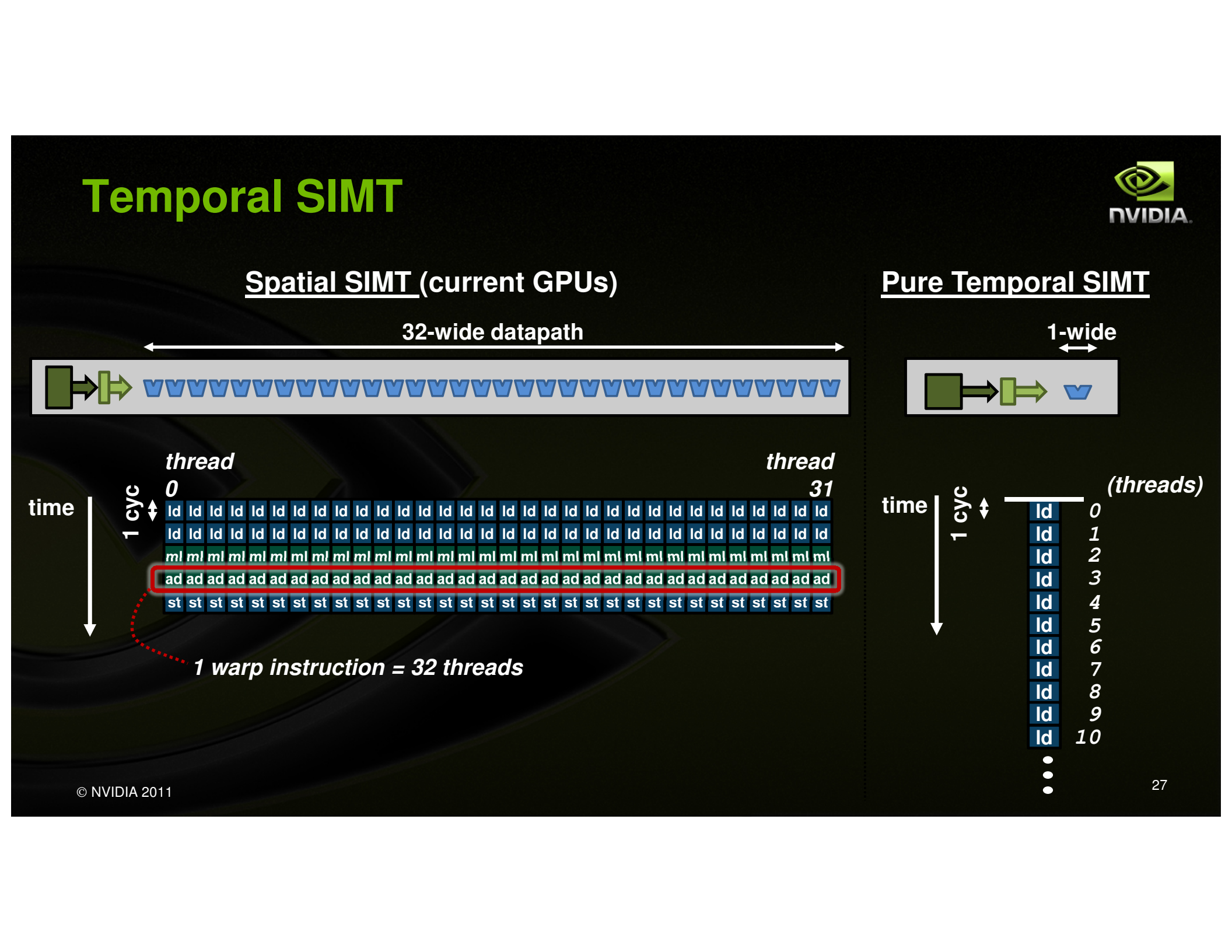
- **SIMD: Single Instruction Multiple Data**
- **MIMD: Multiple Instruction Multiple Data**
- **SIMT: Single Instruction Multiple Thread**



SIMT = MIMD Programming Model with SIMD Implementation Efficiencies

© NVIDIA 2011

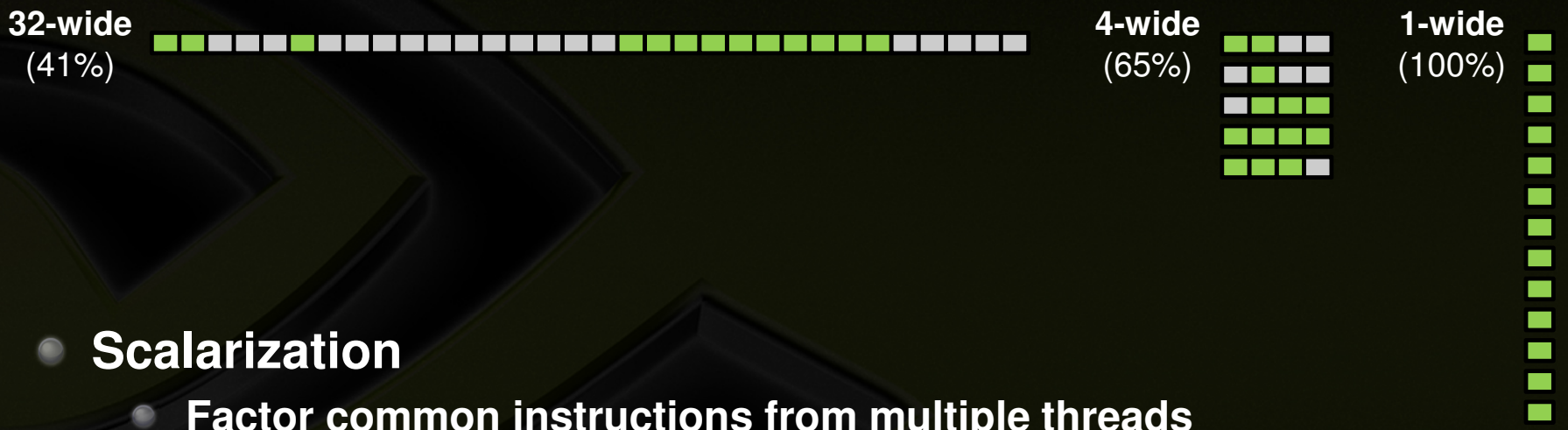
Temporal SIMT



© NVIDIA 2011

Temporal SIMT Optimizations

- Control divergence – hybrid MIMD/SIMT



- Scalarization
  - Factor common instructions from multiple threads
  - Execute once – place results in common registers

© NVIDIA 2011

Strategies for Energy Reduction

- Simpler processor architectures
- Reduce waste
- Improve physical locality of data
- **Exploit heterogeneity and specialization**
- **Signaling and packaging**
- **Inch down voltage**

© NVIDIA 2011

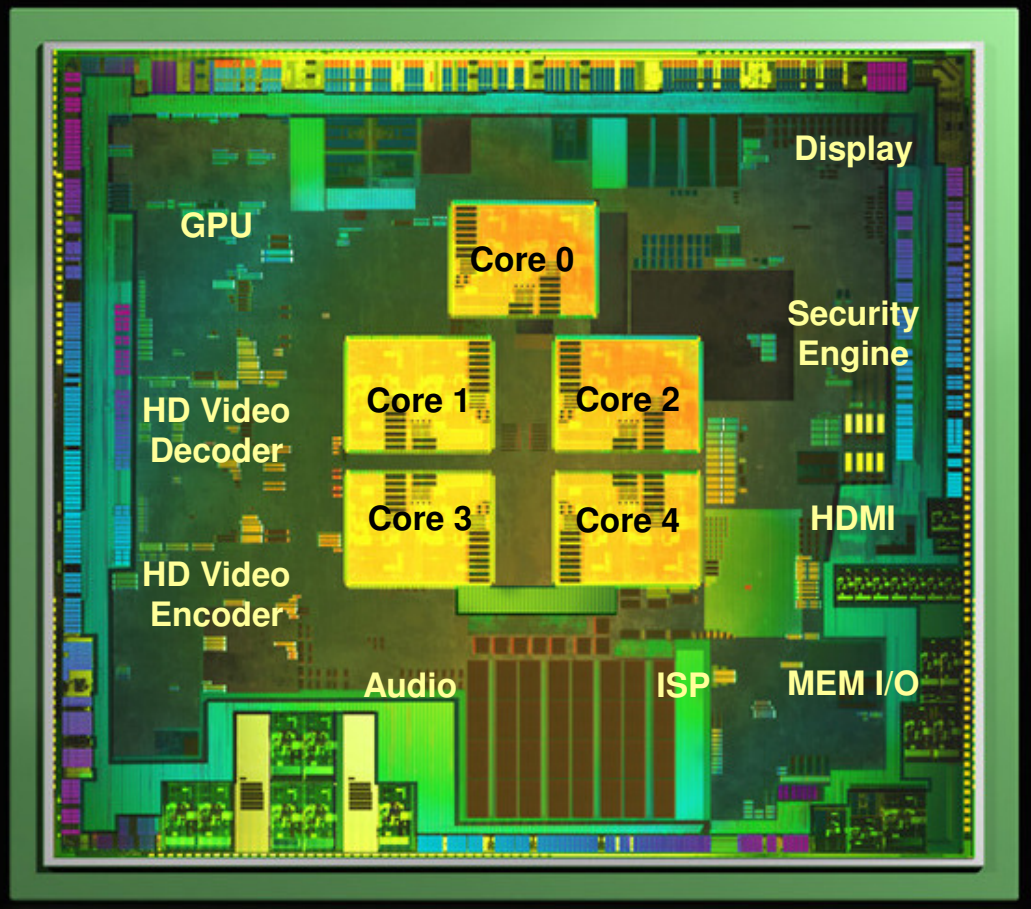
Heterogeneity and Specialization

- Programmable processor heterogeneity
  - CPU + GPU in same system/chip
- Functional specialization
  - Fixed/limited programmability accelerators
- Power specialization
  - Same function, different power/performance profile

© NVIDIA 2011

NVIDIA Tegra3 Mobile SOC

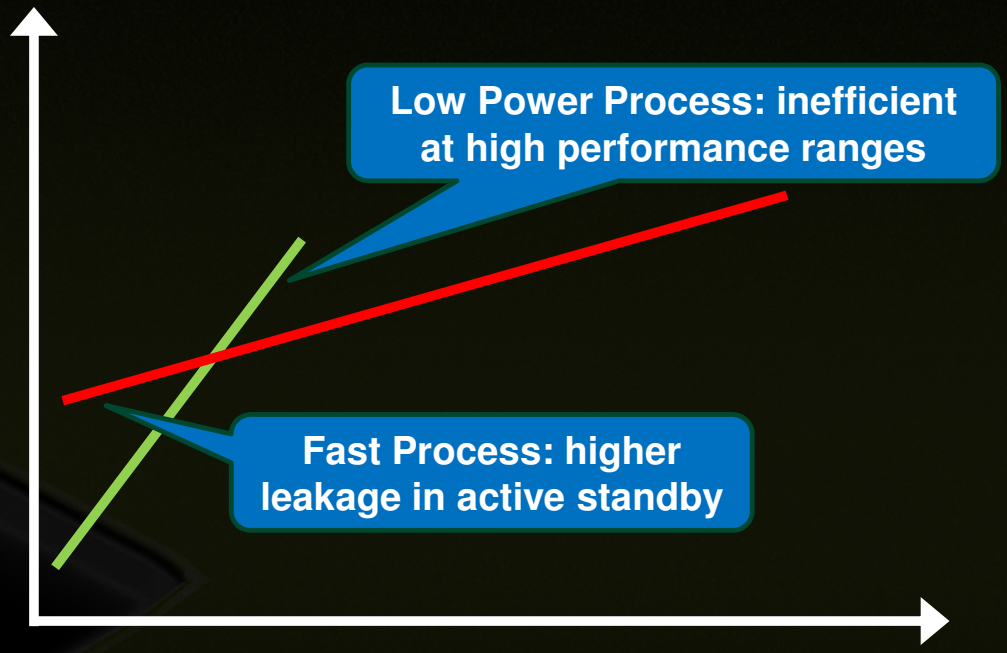
- 5 ARM Cortex A9 Cores
- Range of accelerators
  - Some programmable
  - Some fixed function
- 1-5W power envelope



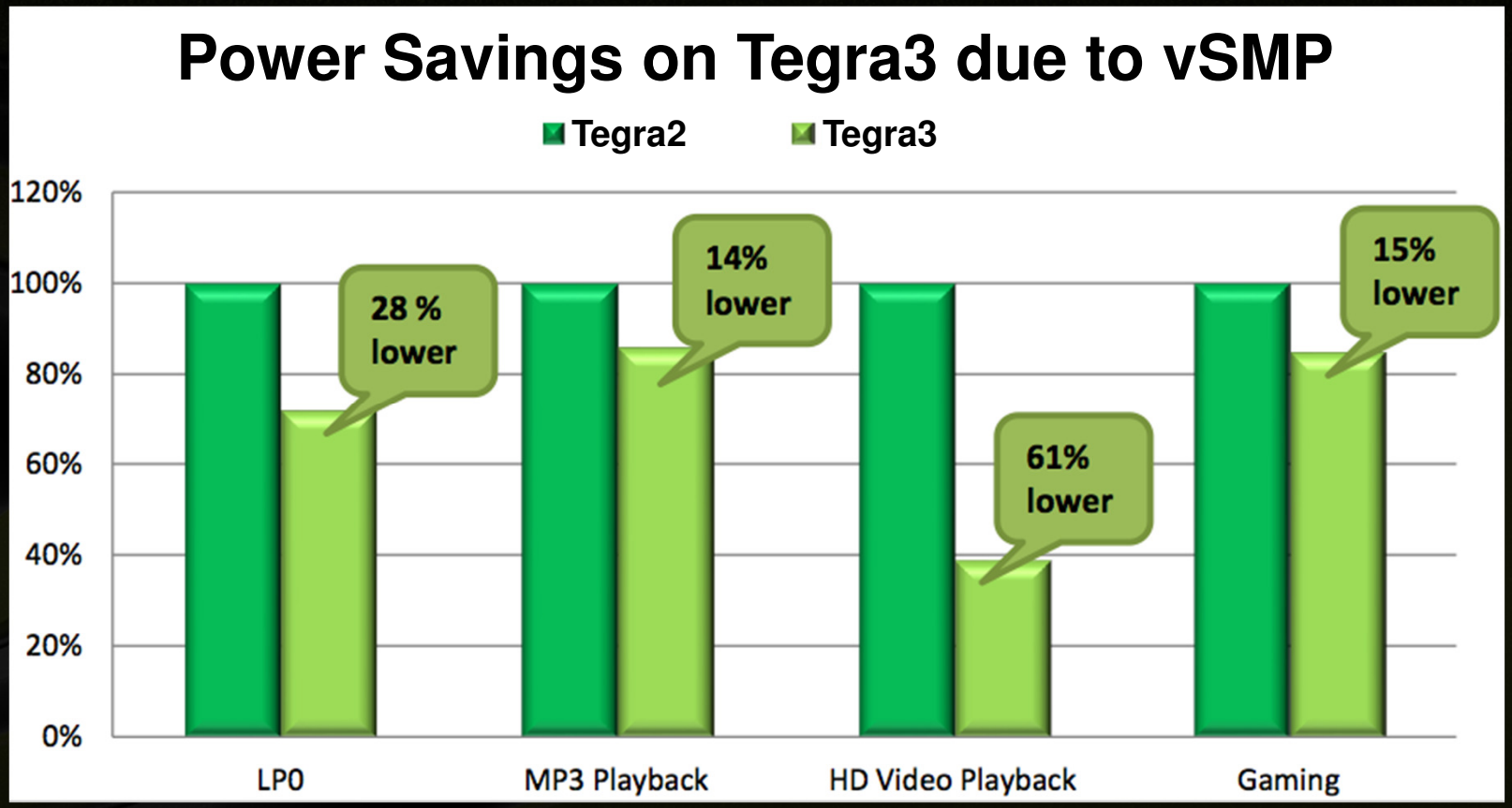
© NVIDIA 2011

Variable SMP

- Same CPU core design
- Optimized for different voltage/frequency range
  - Different transistors
  - Different synthesis

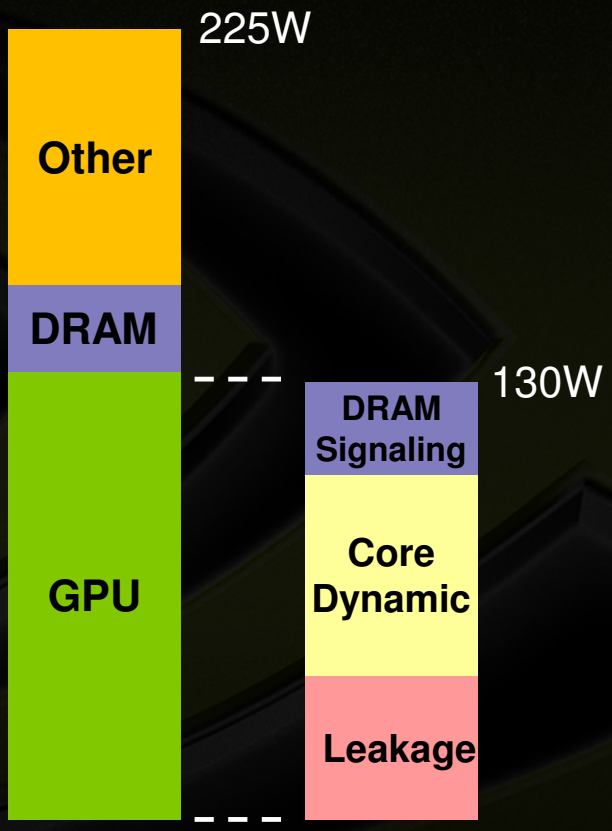


© NVIDIA 2011


Power Savings on Tegra3 due to vSMP
Tegra2
Tegra3
120%
100%
80%
60%
40%
20%
0%
28 % lower
14% lower
61% lower
15% lower
LP0
MP3 Playback
HD Video Playback
Gaming

© NVIDIA 2011

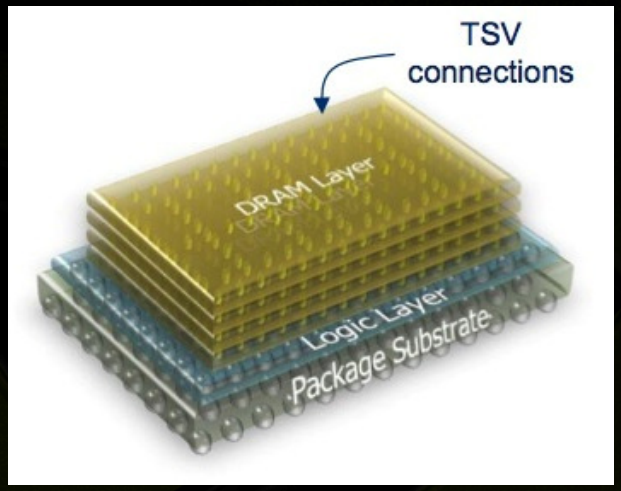
Typical Power Breakdown



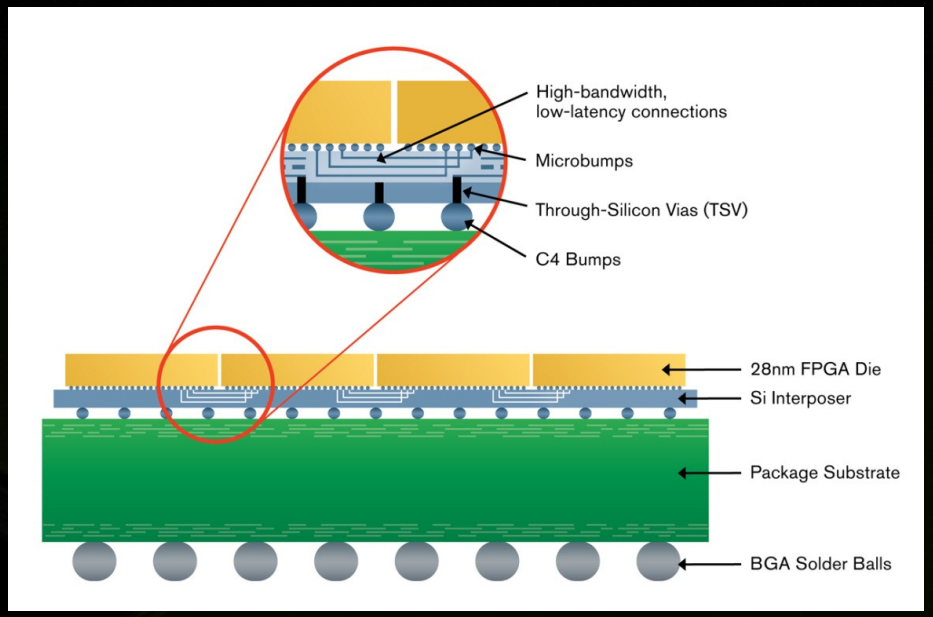
NVIDIA M2070 Module

© NVIDIA 2011

Packaging/Signaling/Architecture



Micron Hybrid Memory Cube



Silicon Interposer Packaging (Xilinx)

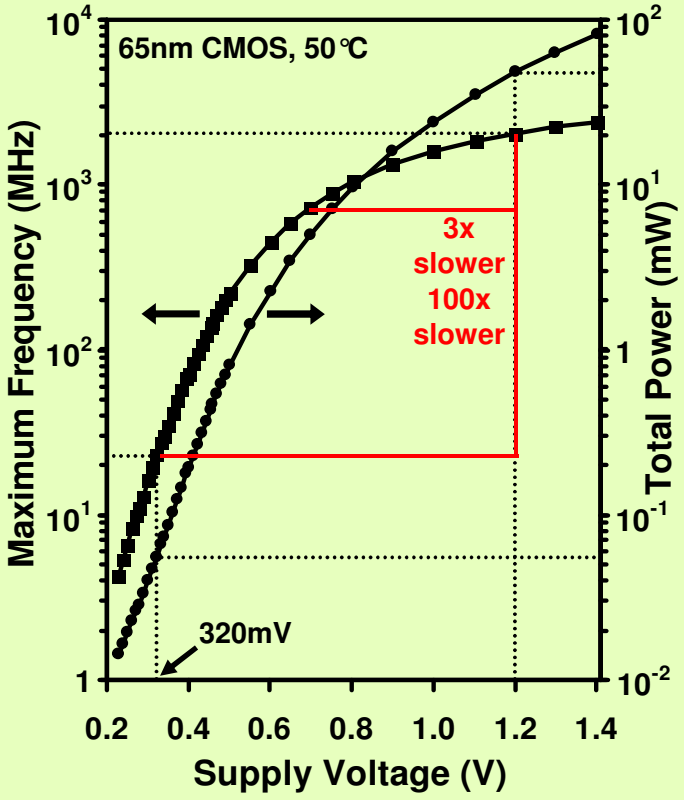
© NVIDIA 2011

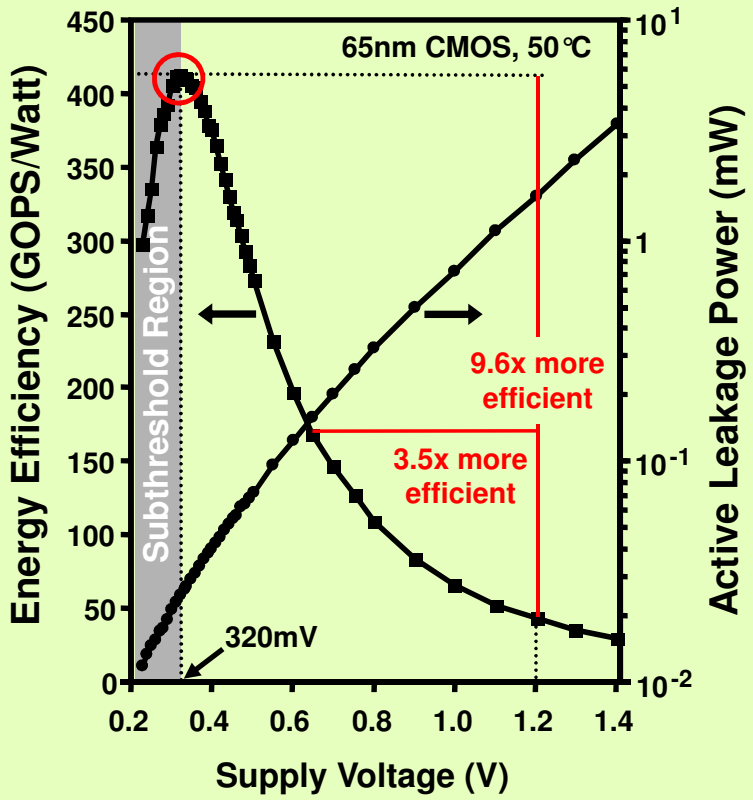
Voltage Scaling – is there any left?

- Answer #1: Hard – scale V_{dd} down toward $2*V_t$
- Answer #2: Harder – scale down to near V_t
- Principle challenges
  - Slow transistors
  - Reliability
  - Process variation

© NVIDIA 2011

Aggressive Voltage Scaling





H. Kaul et al. ISSCC08; Motion Estimator Accelerator

© NVIDIA 2011

Energy Efficiency Potential at 10nm

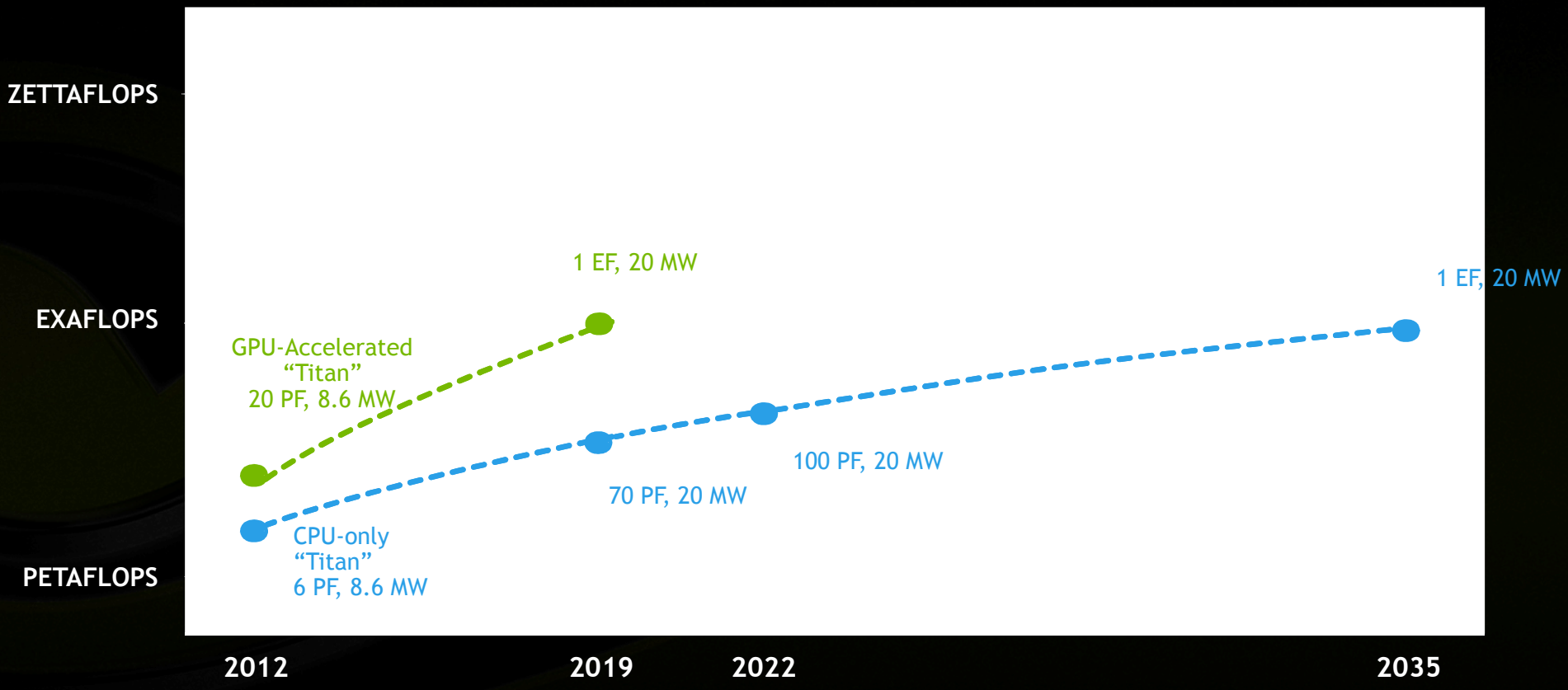
Natural process scaling	6x	Assuming 0.9=>0.7V

Energy efficient chip architecture	3-5x	
Circuits/Packaging	2-2.5x	Includes on-chip ckts
Slower clock frequency	2x	
Aggressive voltage scaling	1.5x	
Total:	20-40x	~3x area penalty

Architecture specialization	~10x	On compute limited applications
Near-threshold Voltage	4x (theoretical)	~30x area penalty

© NVIDIA 2011

The Road to Exascale



© NVIDIA 2011

“Teraflops”

5 Watts

ASCI Red @ Sandia Labs

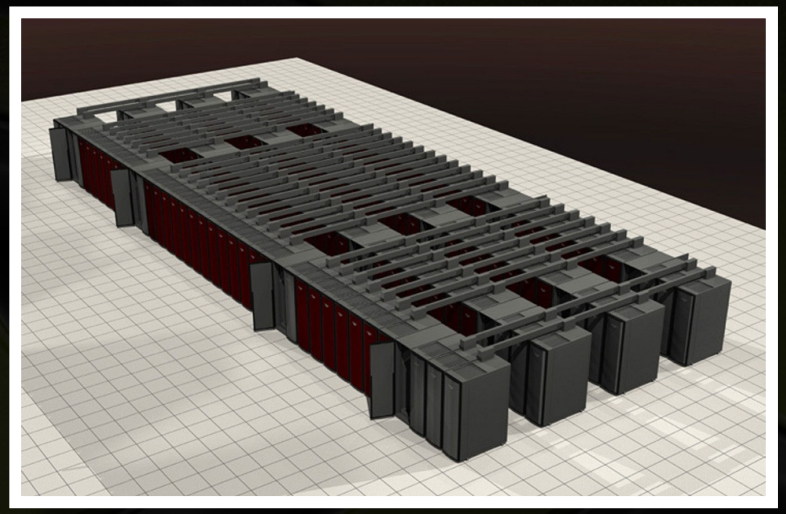
1997

2019

© NVIDIA 2011

"Tens" of Teraflops

100 Watts



Red Storm @ Sandia Labs

2004

2019

© NVIDIA 2011

"Hundreds" of Teraflops

1000 Watts

Blue Gene @ LLNL

2006

2019

© NVIDIA 2011

Challenges for Research Community

- Fundamentally energy efficient architectures
  - Core, memory system, interconnect, system
- Low voltage
  - Clocking, RAM, signaling, variation tolerant
- Specialization
  - In general purpose system – what are the “accelerators”?
- Emerging and important workloads are throughput

© NVIDIA 2011

‘Super’ Computing

From Super Computers to Super Phones