

A detailed, high-angle photograph of an NVIDIA GPU chip mounted on a printed circuit board (PCB). The chip is a large, square, black component with a silver-colored metal heat spreader on top. The NVIDIA logo is prominently displayed in the center of the chip. The PCB is dark green with intricate circuit patterns, various capacitors, and other electronic components. Several circular mounting holes with copper plating are visible around the chip. The lighting is dramatic, highlighting the textures of the chip and the board.

GPU TECHNOLOGY
CONFERENCE

INSIDE PASCAL

Lars Nyland and Mark Harris, April 5, 2016

INTRODUCING TESLA P100

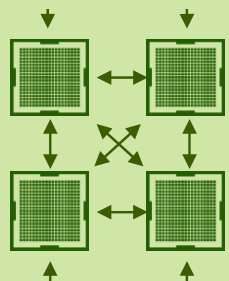
New GPU Architecture to Enable the World's Fastest Compute Node

Pascal Architecture



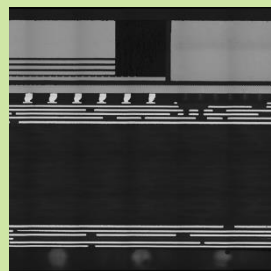
Highest Compute Performance

NVLink



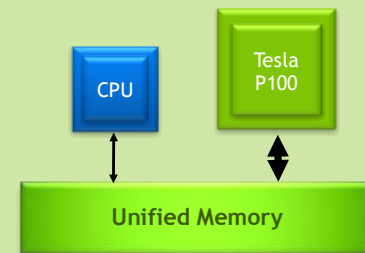
GPU Interconnect for Maximum Scalability

HBM2 Stacked Memory

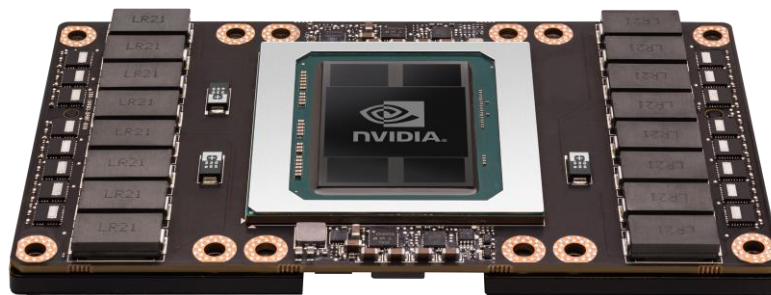


Unifying Compute & Memory in Single Package

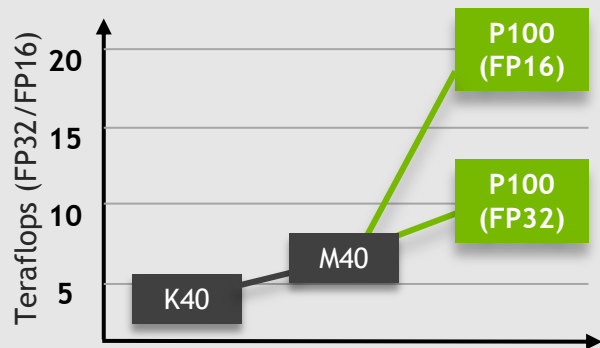
Page Migration Engine



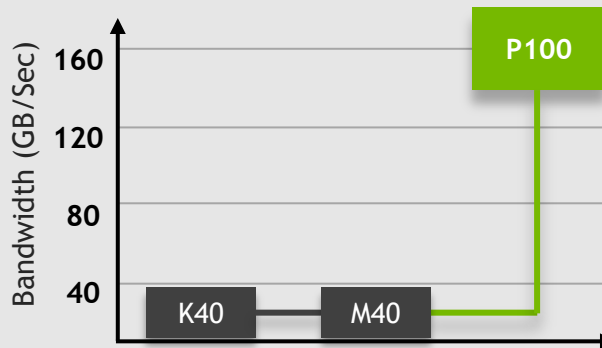
Simple Parallel Programming with 512 TB of Virtual Memory



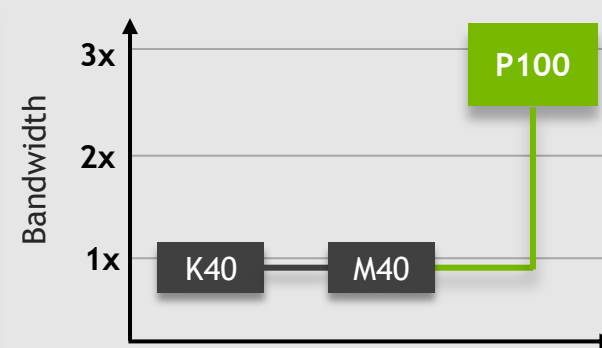
GIANT LEAPS IN EVERYTHING



3x Compute



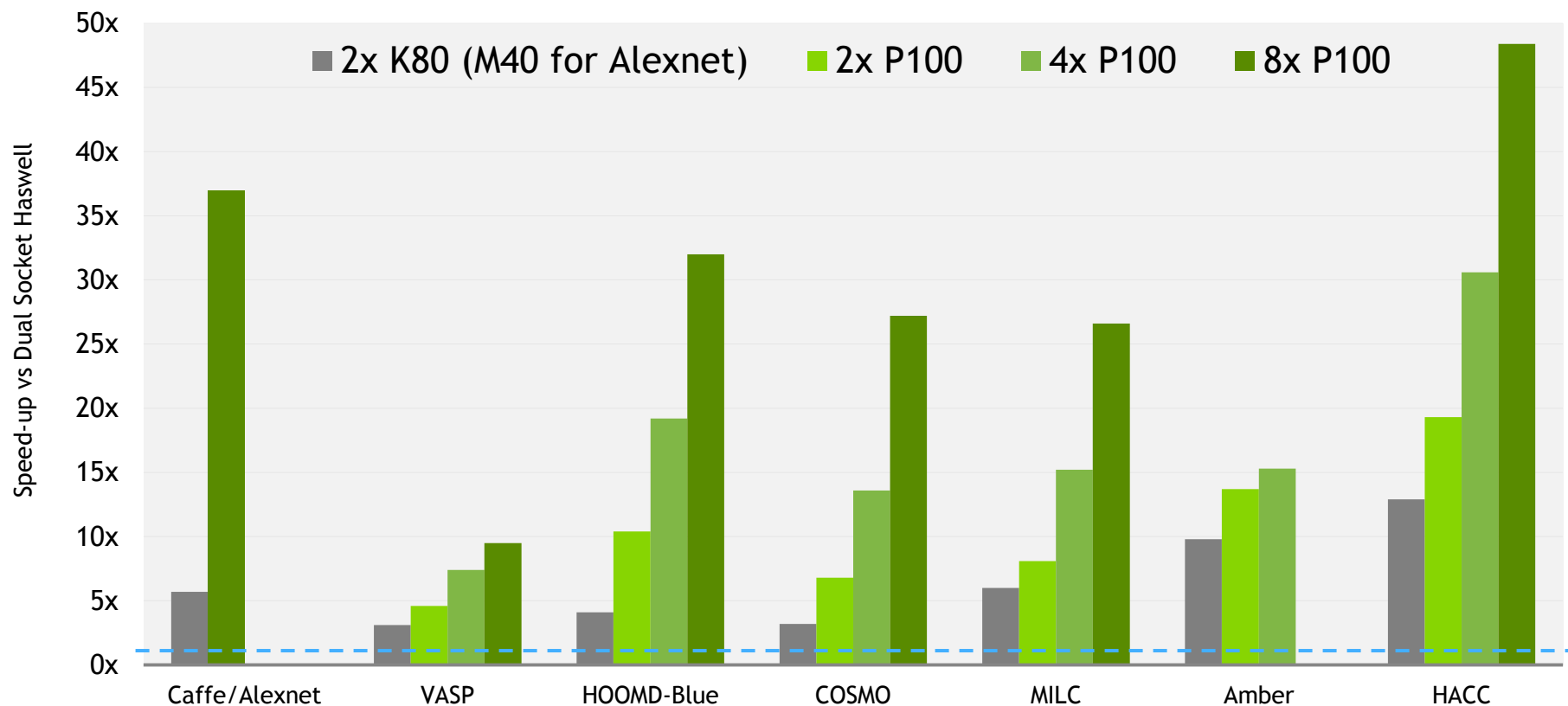
5x GPU-GPU BW



3x GPU Mem BW

TESLA P100 PERFORMANCE DELIVERED

NVLink for Max Scalability, More than 45x Faster with 8x P100



2x Haswell
CPU

PASCAL ARCHITECTURE

TESLA P100 GPU: GP100

56 SMS

3584 CUDA Cores

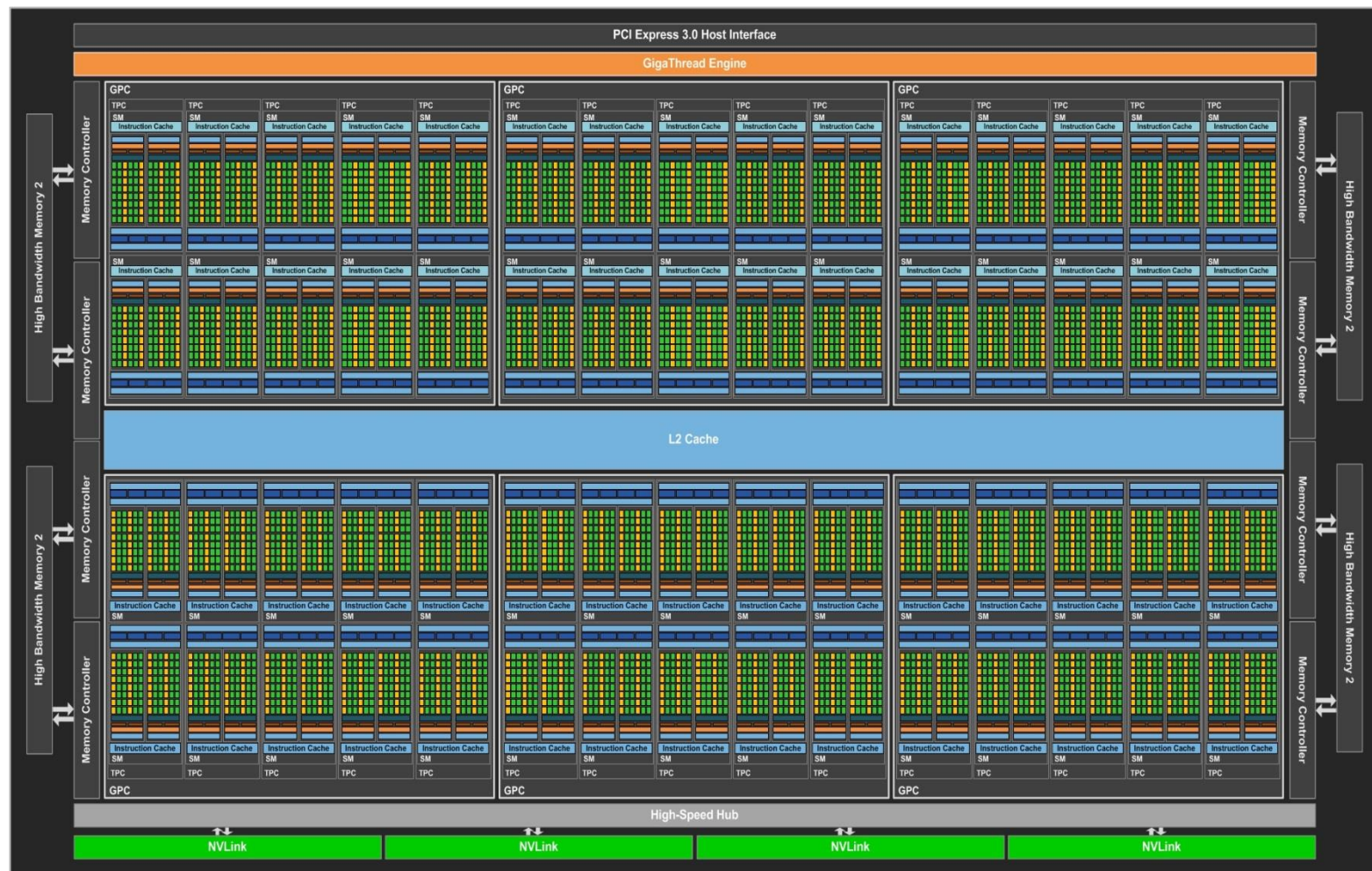
5.3 TF Double Precision

10.6 TF Single Precision

21.2 TF Half Precision

16 GB HBM2

720 GB/s Bandwidth



GPU PERFORMANCE COMPARISON

	P100	M40	K40
Double Precision TFlop/s	5.3	0.2	1.4
Single Precision TFlop/s	10.6	7.0	4.3
Half Precision Tflop/s	21.2	NA	NA
Memory Bandwidth (GB/s)	720	288	288
Memory Size	16GB	12GB, 24GB	12GB

GP100 SM

GP100

CUDA Cores 64

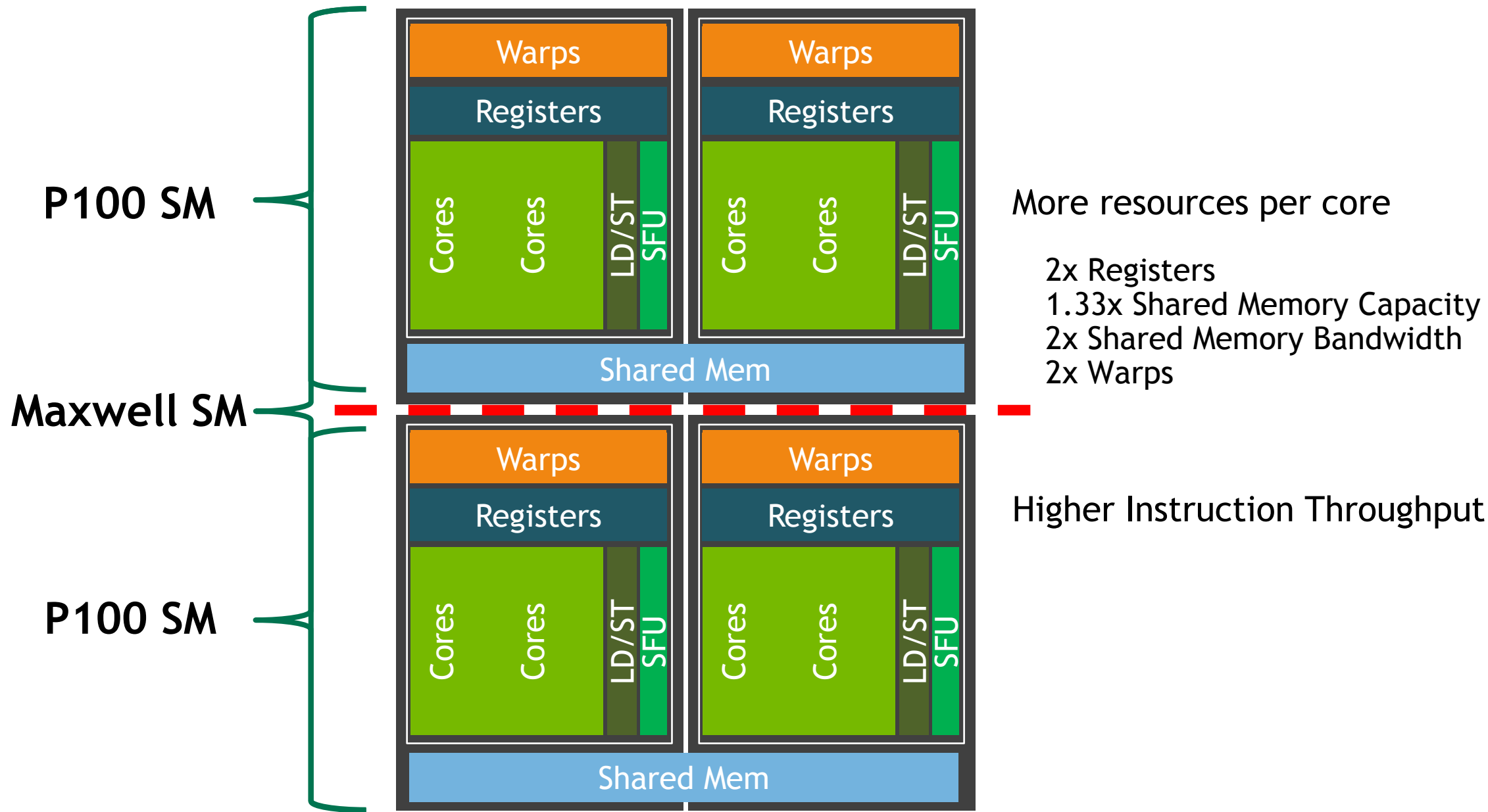
Register File 256 KB

Shared Memory 64 KB

Active Threads 2048



Active Blocks 32





IEEE 754 FLOATING POINT ON GP100

3 sizes, 3 speeds, all fast

Feature	 Half precision	Single precision	Double precision
Layout	s5.10	s8.23	s11.52
Issue rate	pair every clock	1 every clock	1 every 2 clocks
Subnormal support	Yes	Yes	Yes
Atomic Addition	Yes	Yes	 Yes

HALF-PRECISION FLOATING POINT (FP16)

- 16 bits

s	e	x	p			.	f	r	a	c						
---	---	---	---	--	--	---	---	---	---	---	--	--	--	--	--	--

 - 1 sign bit, 5 exponent bits, 10 fraction bits
- 2^{40} Dynamic range
 - Normalized values: 1024 values for each power of 2, from 2^{-14} to 2^{15}
 - Subnormals at full speed: 1024 values from 2^{-24} to 2^{-15}
- Special values
 - +- Infinity, Not-a-number

USE CASES

Deep Learning Training

Radio Astronomy

Sensor Data

Image Processing

NVLink

NVLINK

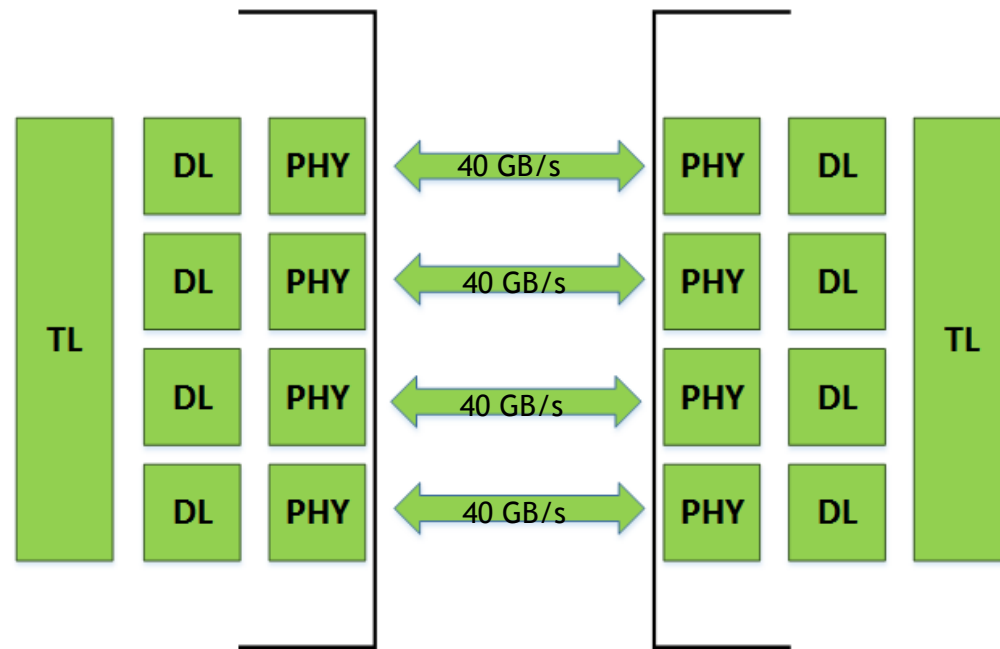
P100 supports 4 NVLinks

Up to 94% bandwidth efficiency

Supports read/writes/atomics to peer GPU

Supports read/write access to NVLink-enabled CPU

Links can be ganged for higher bandwidth



NVLink on Tesla P100

NVLINK - GPU CLUSTER

Two fully connected quads,
connected at corners

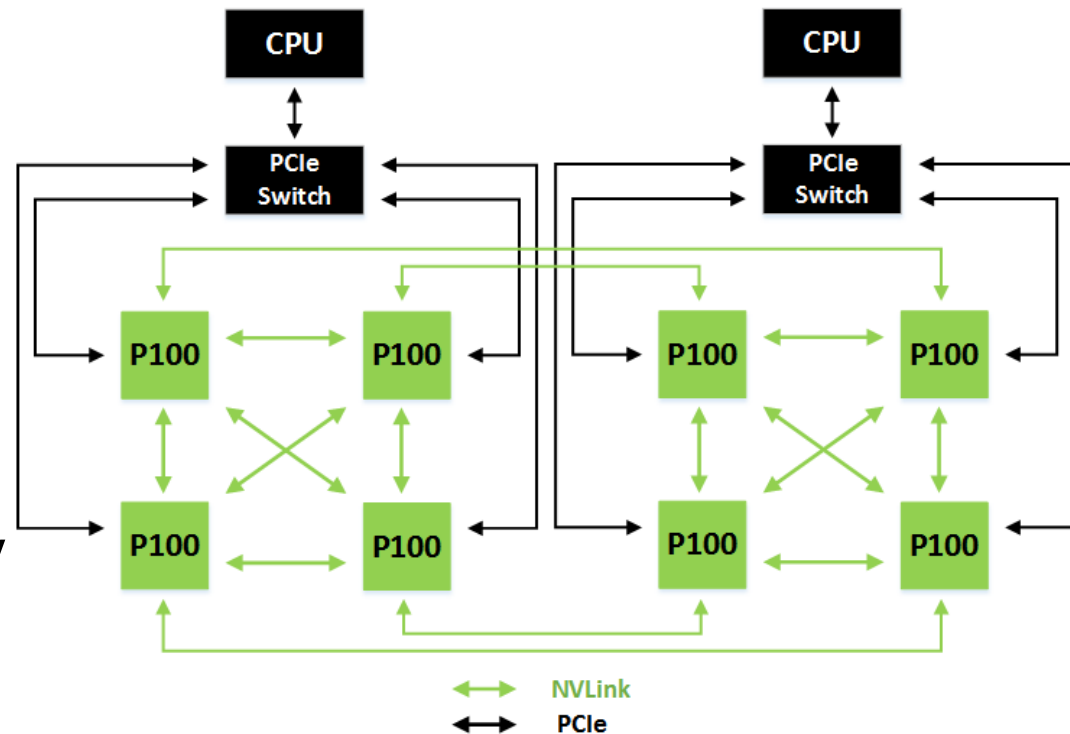
160GB/s per GPU bidirectional to Peers

Load/store access to Peer Memory

Full atomics to Peer GPUs

High speed copy engines for bulk data copy

PCIe to/from CPU



NVLINK TO CPU

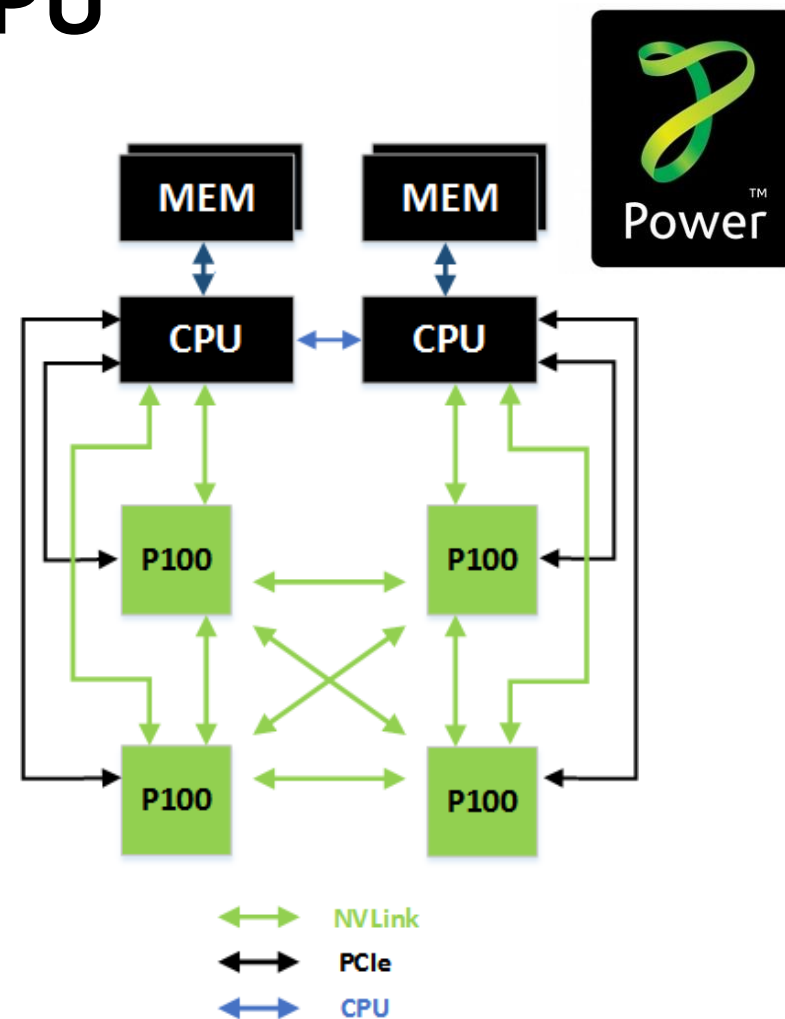
Fully connected quad

120 GB/s per GPU bidirectional for peer traffic

40 GB/s per GPU bidirectional to CPU

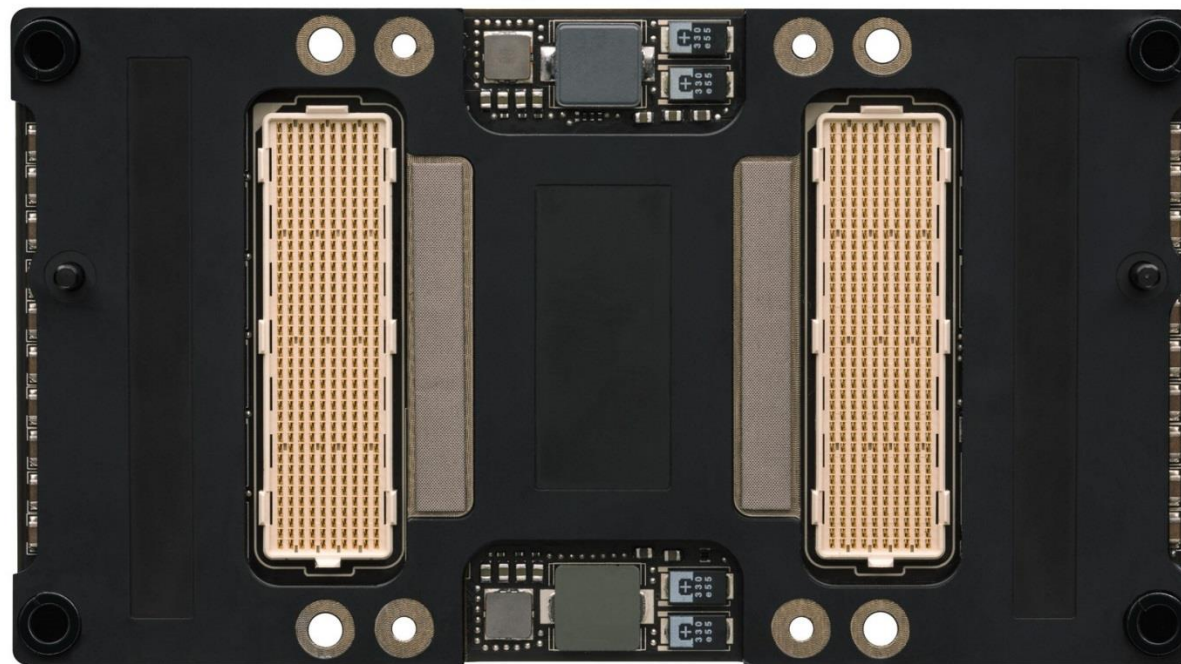
Direct Load/store access to CPU Memory

High Speed Copy Engines for bulk data movement



TESLA P100 PHYSICAL CONNECTOR

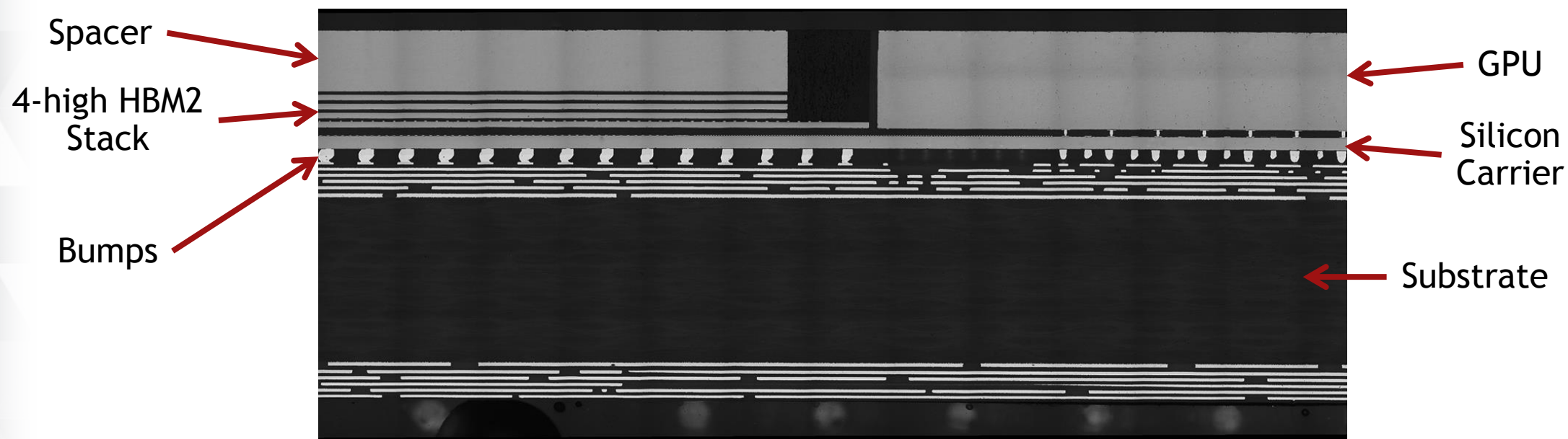
With NVLink



HBM2 STACKED MEMORY

HBM2 : 720GB/SEC BANDWIDTH

And ECC is free



UNIFIED MEMORY

PAGE MIGRATION ENGINE

Support Virtual Memory Demand Paging

49-bit Virtual Addresses

Sufficient to cover 48-bit CPU address + all GPU memory

GPU page faulting capability

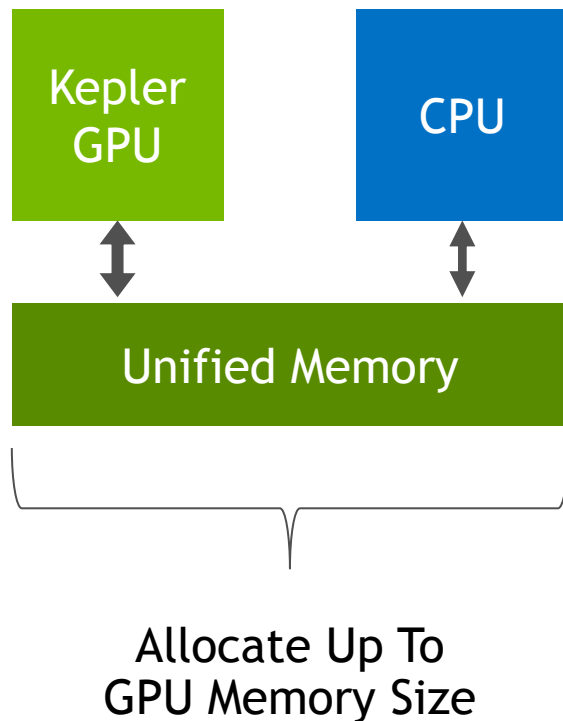
Can handle thousands of simultaneous page faults

Up to 2 MB page size

Better TLB coverage of GPU memory

KEPLER/MAXWELL UNIFIED MEMORY

CUDA 6+



Simpler
Programming &
Memory Model

Single allocation, single pointer,
accessible anywhere
Eliminate need for *explicit copy*
Greatly simplifies code porting

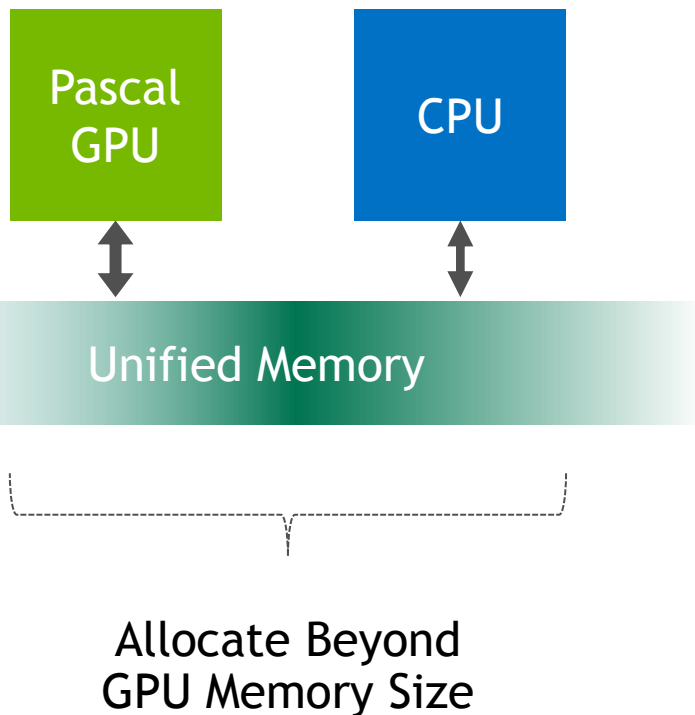
Performance
Through
Data Locality

Migrate data to accessing processor
Guarantee global coherency
Still allows explicit hand tuning

PASCAL UNIFIED MEMORY

Large datasets, simple programming, High Performance

CUDA 8



Enable Large
Data Models

Oversubscribe GPU memory
Allocate up to system memory size

Tune
Unified Memory
Performance

Usage hints via `cudaMemAdvise` API
Explicit prefetching API

Simpler
Data Access

CPU/GPU Data coherence
Unified memory atomic operations

INTRODUCING TESLA P100

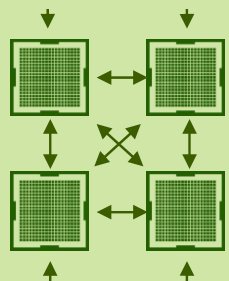
New GPU Architecture to Enable the World's Fastest Compute Node

Pascal Architecture



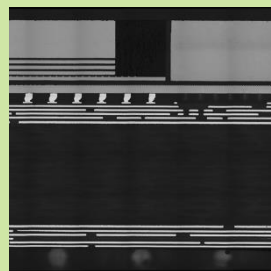
Highest Compute Performance

NVLink



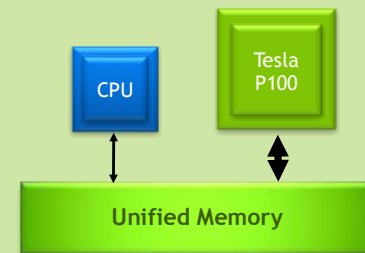
GPU Interconnect for Maximum Scalability

HBM2 Stacked Memory



Unifying Compute & Memory in Single Package

Page Migration Engine



Simple Parallel Programming with 512 TB of Virtual Memory

More P100 Features: compute preemption, new instructions, larger L2 cache, more...

Find out more at <http://devblogs.nvidia.com/parallelforall/inside-pascal>