# NVIDIA VOLTA ARCHITECTURE

Jeff Larkin, NVIDIA December 03, 2018

# **TESLA V100**



#### The Fastest and Most Productive GPU for Deep Learning and HPC

# **TESLA V100**

21B transistors 815 mm<sup>2</sup>

80 SM 5120 CUDA Cores 640 Tensor Cores

16 GB HBM2 900 GB/s HBM2 300 GB/s NVLink



\*full GV100 chip contains 84 SMs

# **GPU PERFORMANCE COMPARISON**

|                        | P100        | V100          | Ratio |
|------------------------|-------------|---------------|-------|
| Training acceleration  | 10 TOPS     | 120 TOPS      | 12x   |
| Inference acceleration | 21 TFLOPS   | 120 TOPS      | 6x    |
| FP64/FP32              | 5/10 TFLOPS | 7.5/15 TFLOPS | 1.5x  |
| HBM2 Bandwidth         | 720 GB/s    | 900 GB/s      | 1.2x  |
| NVLink Bandwidth       | 160 GB/s    | 300 GB/s      | 1.9x  |
| L2 Cache               | 4 MB        | 6 MB          | 1.5x  |
| L1 Caches              | 1.3 MB      | 10 MB         | 7.7x  |

### **NEW HBM2 MEMORY ARCHITECTURE**

1.5x Delivered



V100 measured on pre-production hardware.

# **VOLTA NVLINK**

300GB/sec 50% more links 28% faster signaling



### PROGRAMMABILITY

# PASCAL UNIFIED MEMORY





# **VOLTA + PCIE CPU UNIFIED MEMORY**





# **VOLTA + NVLINK CPU UNIFIED MEMORY**





# **VOLTA MULTI-PROCESS SERVICE**

#### Volta MPS Enhancements:

- Reduced launch latency
- Improved launch throughput
- Improved quality of service with scheduler partitioning
  - More reliable performance
- 3x more clients than Pascal



### NEW SM MICROARCHITECTURE

# VOLTA GV100 SM

|                             | GV100  |
|-----------------------------|--------|
| FP32 units                  | 64     |
| FP64 units                  | 32     |
| INT32 units                 | 64     |
| Tensor Cores                | 8      |
| <b>Register File</b>        | 256 KB |
| Unified L1/Shared<br>memory | 128 KB |
| Active Threads              | 2048   |

| И   |  |   |  |  |  |   |                       |                                       |   |   |   |   |  |  |                                    |                |  |
|---|--|---|--|--|--|---|-----------------------|---------------------------------------|---|---|---|---|--|--|------------------------------------|----------------|--|
| L1 Instruction Cache  |  |   |  |  |  |   |                       |                                       |   |   |   |   |  |  |                                    |                |  |
|   |  | L0 li   | nstruc   | tion C   | ache                                   |   |                       |                                       | L0 Instruction Cache  |   |   |   |  |  |                                    |                |  |
| Warp Scheduler (32 thread/clk)  |  |   |  |  |  |   |                       |                                       | Warp Scheduler (32 thread/clk)  |   |   |   |  |  |                                    |                |  |
| Dispatch Unit (32 thread/clk)   |  |   |  |  |  |   |                       | Dispatch Unit (32 thread/clk)         |   |   |   |   |  |  |                                    |                |  |
|   | Reg  | ister   | File (   | 16,384   | 4 x 32                                 | !-bit)  |                       |                                       |   | Reg   | ister   | File (1   | 16,384   | 1 x 32                                 | -bit)                              |                |  |
| FP64  | INT  | INT   | FP32   | FP32   |  |   |                       | FP                                    | 64  | INT   | INT   | FP32  | FP32   |  |                                    |                |  |
| FP64  | INT  | INT   | FP32   | FP32   |  |   |                       | FP                                    | 64  | INT   | INT   | FP32  | FP32   |  |                                    |                |  |
| FP64  | INT  | INT   | FP32   | FP32   |  |   | TENSOR                | FP                                    | 64  | INT   | INT   | FP32  | FP32   |  |                                    |                |  |
| FP64  | INT  | INT   | FP32   | FP32   | TEN                                    | SOR   |                       | FP                                    | 64  | INT   | INT   | FP32  | FP32   | TEN                                    | SOR                                | TENSOR         |  |
| FP64  | INT  | INT   | FP32   | FP32   | cc                                     | RE  | CORE                  | FP                                    | 64  | INT   | INT   | FP32  | FP32   | 2 CORE                                 | CORE                               |                |  |
| FP64  | INT  | INT   | FP32   | FP32   |  |   |                       | FP                                    | 64  | INT   | INT   | FP32  | FP32   |  |                                    |                |  |
| FP64  | INT  | INT   | FP32   | FP32   |  |   |                       | FP                                    | 64  | INT   | INT   | FP32  | FP32   | $\square$                              |                                    |                |  |
| FP64  | INT  | INT   | FP32   | FP32   | H                                      |   |                       | FP                                    | 64  | INT   | INT   | FP32  | FP32   | H                                      |                                    |                |  |
| LD/ LD/<br>ST ST  | LD/<br>ST  | LD/<br>ST   | LD/<br>ST  | LD/<br>ST  | LD/<br>ST                              | LD/<br>ST   | SFU                   | LD/<br>ST                             | LD/<br>ST   | LD/<br>ST   | LD/<br>ST   | LD/<br>ST   | LD/<br>ST  | LD/<br>ST                              | LD/<br>ST                          | SFU            |  |
| L0 Instruction Cache  |  |   |  |  |  |   |                       |                                       |   |   |   |   |  |  |                                    |                |  |
|   |  | L0 li   | nstruc   | tion C   | ache                                   |   |                       |                                       |   |   | L0 Ir   | nstruc  | tion C   | ache                                   |                                    |                |  |
|   | War  | L0 lı<br>p Sch  | nstruc<br>nedule   | tion C<br>r (32 t  | ache<br>hread                          | /clk)   |                       |                                       |   | War   | L0 Ir<br>p Sch  | nstruct<br>edule  | tion C<br>r (32 tl   | ache<br>hread/                         | clk)                               |                |  |
|   | War<br>Dis   | L0 lı<br>p Sch<br>spatcl  | nstruc<br>nedule<br>h Unit   | tion C<br>r (32 t<br>(32 th  | ache<br>hread<br>read/c                | /clk)<br>:lk)   |                       |                                       |   | War<br>Dis  | L0 Ir<br>p Sch<br>spatch  | edule<br>Unit   | tion C<br>r (32 tl<br>(32 th   | ache<br>hread/<br>read/c               | 'cik)<br>:ik)                      |                |  |
|   | War<br>Dis<br>Reg  | L0 II<br>p Sch<br>spatcl<br>ister   | nstruc<br>nedule<br>h Unit<br>File ('  | tion C<br>r (32 t<br>(32 th<br>16,38   | ache<br>hread<br>read/d<br>4 x 32      | /clk)<br>clk)<br>2-bit)   |                       |                                       |   | War<br>Dis<br>Reg   | L0 Ir<br>p Sch<br>spatch<br>ister   | edulei<br>n Unit<br>File (1   | tion C<br>r (32 tl<br>(32 th<br>16,384   | ache<br>hread/<br>read/c<br>1 x 32     | 'clk)<br>ilk)<br>-bit)             |                |  |
| FP64  | War<br>Di:<br>Reg  | L0 II<br>p Sch<br>spatcl<br>ister<br>INT  | nstruc<br>nedule<br>h Unit<br>File ('<br>FP32  | tion C<br>r (32 t<br>(32 th<br>16,38<br>FP32   | ache<br>hread<br>read/d<br>4 x 32      | /clk)<br>:lk)<br>!-bit)   |                       |                                       | 64  | War<br>Dis<br>Reg   | L0 Ir<br>p Sch<br>spatch<br>ister   | edule<br>n Unit<br>File (1  | tion C<br>r (32 tl<br>(32 th<br>16,384<br>FP32   | ache<br>hread/<br>read/c<br>1 x 32     | clk)<br>ilk)<br>-bit)              |                |  |
| FP64<br>FP64  | War<br>Di:<br>Reg<br>INT   | L0 II<br>p Sch<br>spatcl<br>ister<br>INT<br>INT   | nstruc<br>hedule<br>h Unit<br>File ('<br>FP32<br>FP32  | tion C<br>r (32 th<br>(32 th<br>16,38<br>FP32  | ache<br>hread<br>read/c<br>4 x 32      | /clk)<br>2lk)<br>2-bit)   |                       | FP(                                   | 64  | War<br>Dis<br>Reg<br>INT  | L0 Ir<br>p Sch<br>spatch<br>ister<br>INT<br>INT   | edule<br>Unit<br>File (1<br>FP32<br>FP32  | tion C<br>r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32   | ache<br>hread/c<br>1 x 32              | 'clk)<br>:lk)<br>-bit)             |                |  |
| FP64<br>FP64<br>FP64  | War<br>Di:<br>Reg<br>INT<br>INT  | L0 II<br>p Sch<br>spatcl<br>ister<br>INT<br>INT   | nstruc<br>nedule<br>h Unit<br>File ('<br>FP32<br>FP32<br>FP32  | tion C<br>r (32 t<br>(32 th<br>16,38<br>FP32<br>FP32<br>FP32   | ache<br>hread/c<br>4 x 32              | /clk)<br>:lk)<br>?-bit)   |                       | FPI<br>FPI                            | 64<br>64<br>64  | War<br>Dis<br>Reg<br>INT<br>INT   | L0 Ir<br>p Sch<br>spatch<br>ister<br>INT<br>INT   | edule<br>D Unit<br>File (1<br>FP32<br>FP32<br>FP32  | tion C<br>r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32                                       | ache<br>hread/c<br>1 x 32              | clk)<br>:lk)<br>-bit)              |                |  |
| FP64<br>FP64<br>FP64<br>FP64  | War<br>Dis<br>Reg<br>INT<br>INT<br>INT   | L0 In<br>p Sch<br>spatc<br>ister<br>INT<br>INT<br>INT   | nstruc<br>hedule<br>h Unit<br>File (<br>FP32<br>FP32<br>FP32<br>FP32                                 | tion C<br>(32 th<br>(32 th<br>16,38<br>FP32<br>FP32<br>FP32<br>FP32                                  | ache<br>hread<br>read/o<br>4 x 32      | /clk)<br>:lk)<br>!-bit)   | TENSOR                | FPI<br>FPI                            | 64<br>64<br>64<br>64  | War<br>Dis<br>Reg<br>INT<br>INT<br>INT  | L0 Ir<br>p Sch<br>spatch<br>ister<br>INT<br>INT<br>INT  | edule<br>o Unit<br>File (1<br>FP32<br>FP32<br>FP32<br>FP32                                  | tion C<br>r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32                               | ache<br>hread/<br>read/c<br>1 x 32     | 'clk)<br>:lk)<br>-bit)<br>SOR      | TENSOR         |  |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64                                      | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT                                    | L0 In<br>p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT   | nstruc<br>nedule<br>h Unit<br>File ('<br>FP32<br>FP32<br>FP32<br>FP32                                | tion C<br>(32 th<br>(32 th<br>16,38<br>FP32<br>FP32<br>FP32<br>FP32                                  | ache<br>hread<br>4 x 32<br>TEN<br>CC   | /clk)<br>2-bit)<br>SOR<br>DRE   | TENSOR                | FPI<br>FPI                            | 64<br>64<br>64<br>64<br>64  | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT   | LO Ir<br>p Sch<br>spatch<br>ister<br>INT<br>INT<br>INT<br>INT                                   | eduler<br>Unit<br>File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                           | tion C<br>r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                       | ache<br>hread/c<br>1 x 32<br>TEN<br>CO | cik)<br>ik)<br>-bit)<br>SOR        | TENSOR         |  |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64                              | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT                             | L0 II<br>p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT                                    | nstruc<br>nedule<br>h Unit<br>File (<br>FP32<br>FP32<br>FP32<br>FP32                                 | tion C<br>r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                       | ache<br>hread/<br>4 x 32               | /clk)<br>lk)<br>t-bit)<br>SOR   | TENSOR                | FPI<br>FPI<br>FPI                     | 64<br>64<br>64<br>64<br>64<br>64                                      | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT                                    | L0 Ir<br>p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT                            | eduler<br>o Unit<br>File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                 | tion C<br>(32 th<br>(32 th<br>(32 th<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                         | ache<br>hread/<br>read/c<br>1 x 32     | clk)<br>Ik)<br>-bit)<br>SOR<br>RE  | TENSOR         |  |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64                      | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT                             | LO In<br>p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                             | nstruc<br>nedule<br>h Unit<br>File (<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                 | tion C<br>r (32 th<br>(32 th<br>16,38<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32        | ache<br>hread<br>read/c<br>4 x 32      | /clk)<br>slk)<br>t-bit)<br>SOR<br>PRE   | TENSOR                | FPI<br>FPI<br>FPI<br>FPI              | 64<br>64<br>64<br>64<br>64<br>64                                      | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                             | L0 Ir<br>p Sch<br>spatch<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT                            | restruction<br>eduled<br>File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32    | tion C<br>r (32 th<br>(32 th<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                 | ache<br>hread/<br>read/c<br>1 x 32     | clk)<br>lik)<br>-bit)<br>SOR<br>RE | TENSOR         |  |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64              | War<br>Di:<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                      | LO II<br>p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                      | nstruc<br>nedule<br>h Unit<br>File (<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32         | tion C<br>r (32 tt<br>(32 tth<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32      | ache<br>hread<br>4 x 32<br>TEN<br>CC   | /clk)<br>:lk)<br>:-bit)<br>SOR  | TENSOR                | FPI<br>FPI<br>FPI<br>FPI              | 64<br>64<br>64<br>64<br>64<br>64<br>64<br>64                          | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                             | L0 Ir<br>p Sch<br>spatch<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                     | edulee<br>o Unit<br>File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | tion C<br>(32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | ache<br>hread/<br>read/c<br>1 x 32     | clk)<br>lk)<br>-bit)<br>SOR<br>RE  | TENSOR         |  |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64              | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT        | L0 In<br>p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | nstruc<br>neduleduleduleduleduleduleduleduleduledul  | tion C<br>r (32 t<br>(32 th<br>16,38<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | ache<br>hread/o<br>4 x 32<br>TEN<br>CC | /clk)<br>/kb)<br>/kb)<br>/kb)<br>/kb)<br>/kb)<br>/kb)<br>/kb)<br>/k   | TENSOR<br>CORE        | FPI<br>FPI<br>FPI<br>FPI<br>FPI       | 64<br>64<br>64<br>64<br>64<br>64<br>64<br>64<br>64<br>64<br>10/<br>ST | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT        | LO Ir<br>p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>LD/<br>ST | eduler<br>o Unit<br>File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | tion C<br>(32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | ache<br>hread/c<br>1 x 32<br>TEN<br>CO | clk)<br>ilk)<br>-bit)<br>SOR<br>RE | TENSOR<br>CORE |  |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>LDY LDY<br>ST LDY | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | L0 In<br>p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | nstruc<br>nedule<br>h Unit<br>File (<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | tion C<br>r (32 t<br>(32 th<br>16,38<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | ache<br>hread/<br>4 x 32<br>TEN<br>CC  | /clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)<br>/clk)/clk)<br>/clk)/clk)/clk)/clk)/clk)/clk)/clk)/clk) | TENSOR<br>CORE<br>SFU | FPI<br>FPI<br>FPI<br>FPI<br>LD/<br>ST | 64<br>64<br>64<br>64<br>64<br>64<br>64<br>64<br>64<br>1D/<br>ST       | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | LO In<br>p Sch<br>apatol<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>LD/<br>ST | edulei<br>edulei<br>File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | tion C<br>(32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | ache<br>hread/c<br>i x 32<br>TEN<br>CO | clk)<br> lk)<br>-bit)<br>SOR<br>RE | TENSOR<br>CORE |  |

13 💿 nvidia

# VOLTA GV100 SM

**Redesigned for Productivity** 

Completely new ISA

Twice the schedulers

Simplified Issue Logic

Large, fast L1 cache

Improved SIMT model

Tensor acceleration

The easiest SM to program yet

|   |   |  |  |   |   |  | L1 Instruc                      | tion C                         | Cache  |  |   |  |  |  |   |        |
|---|---|--|--|---|---|--|---------------------------------|--------------------------------|--|--|---|--|--|--|---|--------|
|   |   | L0 Ir  | nstruc   | tion C  | ache  |  |                                 |                                |  |  | L0 Ir   | nstruc   | tion C   | ache   |   |        |
| Warp Scheduler (32 thread/clk)  |   |  |  |   |   |  |                                 | Warp Scheduler (32 thread/clk) |  |  |   |  |  |  |   |        |
| Dispatch Unit (32 thread/clk) Dispatch Unit (32 thread/clk)   |   |  |  |   |   |  |                                 |                                |  |  |   |  |  |  |   |        |
| Register File (16,384 x 32-bit)   |   |  |  |   |   |  | Register File (16,384 x 32-bit) |                                |  |  |   |  |  |  |   |        |
| FP64  | INT   | INT  | FP32   | FP32  | $\square$                                     | $\mathbb{H}$   |                                 |                                | FP64   | INT  | INT   | FP32   | FP32   |  |   |        |
| FP64  | INT   | INT  | FP32   | FP32  | ++  |  |                                 |                                | FP64   | INT  | INT   | FP32   | FP32   |  |   |        |
| FP64  | INT   | INT  | FP32   | FP32  | $\blacksquare$                                |  |                                 |                                | FP64   | INT  | INT   | FP32   | FP32   |  |   |        |
| FP64  | INT   | INT  | FP32   | FP32  | TENS  | OR   | TENSOR                          |                                | FP64   |  | INT   | FP32   | FP32   | TENSOR<br>CORE   | TENSOR  |        |
| FP64  | INT   | INT  | FP32   | FP32  | COR   | E  | CORE                            | FP64<br>FP64                   |  | INT  | INT   | FP32   | FP32   |  | CORE  |        |
| FP64  | INT   | INT  | FP32   | FP32  |   |  |                                 |                                |  | INT  | INT   | FP32   | FP32   |  |   |        |
| FP64  | INT   | INT  | FP32   | FP32  |   |  |                                 |                                | FP64   | INT  | INT   | FP32   | FP32   |  |   |        |
| FP64  | INT   | INT  | FP32   | FP32  | $\pm\pm$                                      | $\pm$  |                                 |                                | FP64   | INT  | INT   | FP32   | FP32   | $\vdash$   |   |        |
| LD/ LD/<br>ST ST  | LD/<br>ST   | LD/<br>ST  | LD/<br>ST  | LD/<br>ST   | LD/ L<br>ST S                                 | LD/<br>ST  | SFU                             | L                              | D/ LD/<br>T ST   | LD/<br>ST  | LD/<br>ST   | LD/<br>ST  | LD/<br>ST  | LD/<br>ST  | LD/<br>ST   | SFU    |
| Warp Scheduler (32 thread/clk)     Warp Scheduler (32 thread/clk)       Dispatch Unit (32 thread/clk)     Dispatch Unit (32 thread/clk)       Register File (16 384 x 32-bit)     Register File (16 384 x 32-bit) |   |  |  |   |   |  |                                 |                                |  |  | L0 Ir   | nstruc   | tion C   | ache   |   |        |
|   | War<br>Dis<br>Reg   | p Sch<br>spatcl<br>ister   | edule<br>n Unit<br>File ('   | r (32 tl<br>(32 th<br>16,384  | ache<br>hread/cl<br>read/clk<br>4 x 32-b      | lk)<br>()<br>pit)  |                                 |                                |  | War<br>Dis<br>Reg  | LU Ir<br>p Sch<br>spatch<br>ister   | nstruc<br>ledule<br>n Unit<br>File (1  | tion C<br>r (32 t<br>(32 th<br>16,38 <sup>,</sup>  | ache<br>hread<br>read/c<br>4 x 32  | /clk)<br>clk)<br>2-bit)   |        |
| FDE4  | War<br>Dis<br>Reg   | p Sch<br>spatcl<br>ister   | edule<br>1 Unit<br>File (*   | r (32 tl<br>(32 th<br>16,384  | ache<br>hread/cl<br>read/clk<br>4 x 32-b      | ik)<br>()<br>pit)  |                                 |                                | ED64   | War<br>Dis<br>Reg  | LO Ir<br>p Sch<br>spatch<br>ister   | FILE (1  | tion C<br>r (32 t<br>(32 th<br>16,384  | ache<br>hread<br>read/c<br>4 x 32  | /clk)<br>:lk)<br>?-bit)   |        |
| FP64  | War<br>Dis<br>Reg<br>INT  | p Sch<br>spatcl<br>ister<br>INT  | edule<br>n Unit<br>File (*<br>FP32   | r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32  | ache<br>hread/cl<br>read/clk<br>4 x 32-b      | ik)<br>()<br>pit)  |                                 |                                | FP64   | War<br>Di:<br>Reg<br>INT   | LO Ir<br>p Sch<br>spatch<br>ister<br>INT  | nstruc<br>nedule<br>n Unit<br>File (*<br>FP32  | tion C<br>r (32 t<br>(32 th<br>16,384<br>FP32<br>FP32  | ache<br>hread<br>read/o<br>4 x 32  | /clk)<br>clk)<br>?-bit)   |        |
| FP64<br>FP64<br>FP64  | War<br>Dis<br>Reg<br>INT<br>INT   | p Sch<br>spatcl<br>ister<br>INT<br>INT   | edule<br>1 Unit<br>File (*<br>FP32<br>FP32<br>FP <u>32</u>                                 | r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32  | ache<br>hread/cl<br>read/clk<br>4 x 32-b      | ik)<br>()<br>pit)  |                                 |                                | FP64<br>FP64<br>FP64   | War<br>Dis<br>Reg<br>INT<br>INT  | LU Ir<br>p Sch<br>spatch<br>ister<br>INT<br>INT   | edule<br>n Unit<br>File (*<br>FP32<br>FP32<br>FP32   | tion C<br>r (32 t<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32  | ache<br>Ihread<br>Iread/c<br>4 x 32  | /clk)<br>blk)<br>2-bit)   |        |
| FP64<br>FP64<br>FP64<br>FP64  | War<br>Dis<br>Reg<br>INT<br>INT<br>INT  | p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT                                    | edule<br>Unit<br>File (*<br>FP32<br>FP32<br>FP32<br>FP32                                   | r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32  | ache<br>hread/cl<br>4 x 32-b                  | ik)<br>()<br>pit)  | TENSOR                          |                                | FP64<br>FP64<br>FP64<br>FP64                                 | War<br>Dis<br>Reg<br>INT<br>INT<br>INT   | LUIR<br>p Sch<br>spatch<br>ister<br>INT<br>INT<br>INT                                     | File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | tion C<br>r (32 t<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32                                  | ache<br>hread<br>read/c<br>4 x 32  | /clk)<br>clk)<br>2-bit)   | TENSO  |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64  | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT   | p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT                             | edule<br>Unit<br>File (*<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                           | r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | ache<br>hread/clk<br>4 x 32-b<br>TENS(<br>COR | Ik)<br>()<br>bit)<br>OR  | TENSOR                          |                                | FP64<br>FP64<br>FP64<br>FP64<br>FP64                         | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT                                    | INT<br>INT<br>INT<br>INT<br>INT   | FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32   | tion C<br>r (32 t<br>(32 th<br>16,38<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                           | tene<br>thread<br>read/c<br>4 x 32   | /clk)<br>:lk)<br>2-bit)<br>SOR<br>DRE   | TENSO  |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64  | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT                                  | p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT                      | edule<br>Unit<br>File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                   | r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | ache<br>hread/clk<br>4 x 32-b<br>TENS(<br>COR | lk)<br>k)<br>bit)<br>OR  | TENSOR                          |                                | FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64                 | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT                             | INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT   | rstruction<br>redule<br>n Unit<br>File (*<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | tion C<br>r (32 t<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                  | ache<br>hread<br>read/c<br>4 x 32  | /clk)<br>:lk)<br>2-bit)<br>ISOR<br>DRE  | TENSO  |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64  | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                           | p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT                      | edule<br>1 Unit<br>File ('<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                 | r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | ache<br>hread/cl<br>4 x 32-b<br>TENS(<br>COR  | lk)<br>s)<br>bit)  | TENSOR                          |                                | FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64                 | War<br>Dif<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT                             | ID IF   | Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instru | tion C<br>r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32         | iache<br>hread<br>aread/c<br>4 x 32  | /clk)<br>lbk)<br>P-bit)<br>SOR<br>DRE   | TENSOI |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64  | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                           | p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT               | edule<br>1 Unit<br>File (*<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | <ul> <li>r (32 th</li> <li>(32 th</li> <li>(32 th</li> <li>16,384</li> <li>FP32</li> </ul>  | ache<br>hread/clk<br>4 x 32-b<br>TENS(<br>COR | Ik)<br>(i)<br>bit)<br>OR<br>RE                                     | TENSOR                          |                                | FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64 | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                      | INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                                      | redule<br>edule<br>File (*<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32   | tion C<br>r (32 th<br>(32 th<br>16,38<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | teche<br>thread<br>tread/c<br>tread/c<br>tread/c<br>tread/c<br>tread/c<br>tread/c<br>tread/c<br>tread/c<br>tread/c | /clk)<br>Clk)<br>2-bit)<br>ISOR<br>DRE  | TENSO  |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64  | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>LD/<br>S | p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | edule<br>1 Unit<br>File (*<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | (32 tt)<br>(32 tt)<br>(32 th)<br>(32 th | LD/<br>ST                                     | Ik)<br>k)<br>bit)<br>OR<br>RE<br>LD/<br>ST                         | TENSOR<br>CORE                  |                                | FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64 | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT        | ID IF<br>P Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | redule<br>n Unit<br>File (7<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | tion C<br>r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | LD/<br>ST  | /clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk)<br>clk) | TENSOF |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64  | War<br>Dis<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>LD/             | p Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | edule<br>I Unit<br>File (7<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | tenso<br>TENSO<br>COR                         | (k)<br>(k)<br>(k)<br>(k)<br>(k)<br>(k)<br>(k)<br>(k)<br>(k)<br>(k) | TENSOR<br>CORE<br>SFU           | L S                            | FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64 | War<br>Di:<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | ID IF   | Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instruction<br>Instru | tion C<br>(32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32   | LD/  | /clk)<br>clk)<br>c-bit)<br>SOR<br>DRE<br>LD/<br>ST  | TENSOF |

14 🚳 nvidia

# **RECAP: PASCAL L1 AND SHARED MEMORY**



### **UNIFYING KEY TECHNOLOGIES**



# VOLTA L1 AND SHARED MEMORY

Volta Streaming L1\$ :

Unlimited cache misses in flight Low cache hit latency 4x more bandwidth 5x more capacity

Volta Shared Memory :

Unified storage with L1 Configurable up to 96KB



# NARROWING THE SHARED MEMORY GAP

### with the GV100 L1 cache

Cache: vs shared

- Easier to use
- 90%+ as good

Shared: vs cache

- Faster atomics
- More banks
- More predictable



### **INDEPENDENT THREAD SCHEDULING**

# **VOLTA: INDEPENDENT THREAD SCHEDULING**

#### **Communicating Algorithms**





Pascal: Lock-Free Algorithms

Threads cannot wait for messages

Volta: Starvation Free Algorithms

Threads may wait for messages



# **VOLTA'S EXTENDED SIMT MODEL**

The SIMT model:

enable thread-parallel programs to execute with vector efficiency

|                    | CPU  | Pascal GPU          | Volta GPU |
|--------------------|------|---------------------|-----------|
| Thread-parallelism | MIMD | SIMT<br>(lock-free) | SIMT      |
| Data-parallelism   | SIMD | SIMT                | SIMT      |

# **VOLTA TENSOR CORE**



# **TENSOR CORE**

#### Mixed Precision Matrix Math 4x4 matrices



D = AB + C

# **TENSOR SYNCHRONIZATION**

### Full Warp 16x16 Matrix Math



# **TESLA V100**



More V100 Features: 2x L2 atomics, int8, new memory model, copy engine page migration, and more ...

The Fastest and Most Productive GPU for Deep Learning and HPC