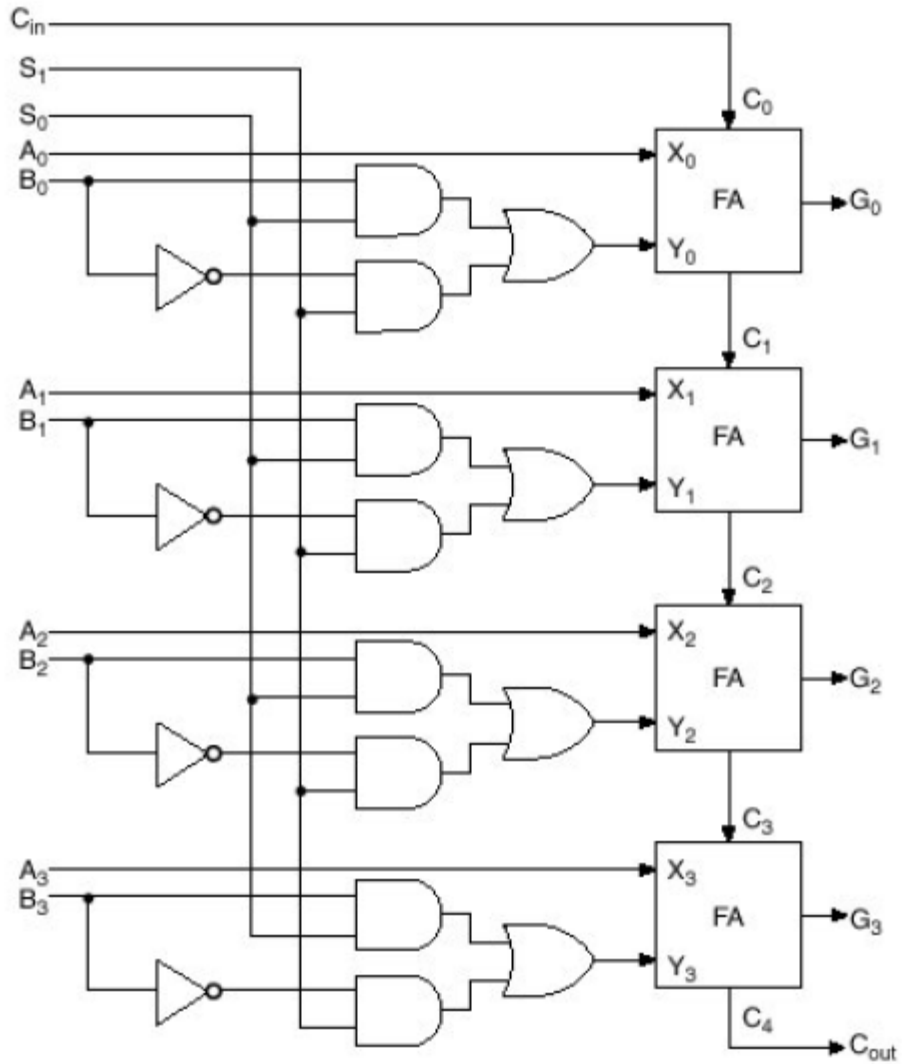
Graphics Hardware

CS7GV3 – Real-time Rendering

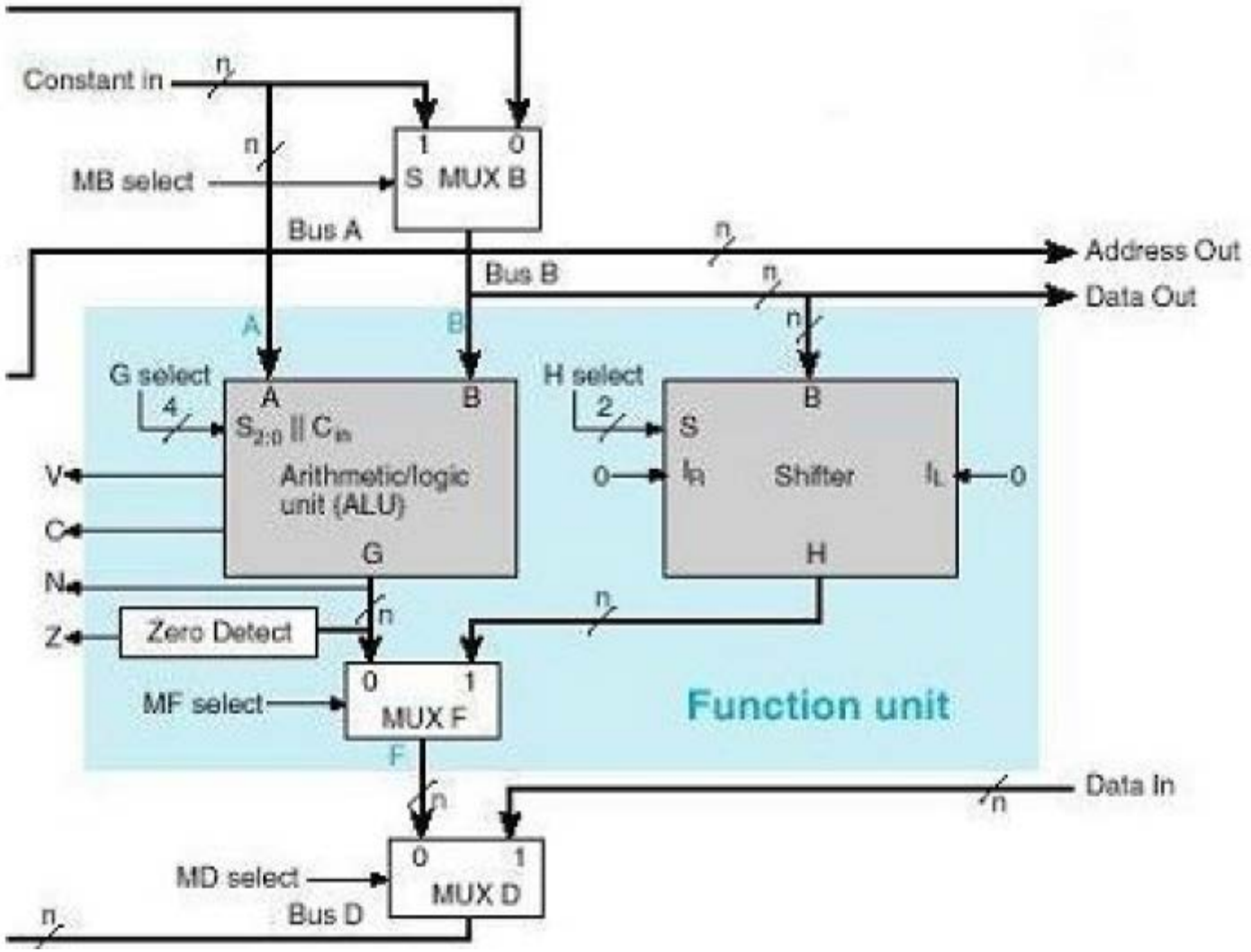
Digital Logic and Arithmetric

- Ripple Carry Adder



Morris Mano and Charles R. Kime. Logic and Computer Design Fundamentals. Prentice-Hall, second edition updated edition, 2001

Function Unit



Morris Mano and Charles R. Kime. Logic and Computer Design Fundamentals. Prentice-Hall, second edition updated edition, 2001

Fundamental Micro-operations

- Register transfer
- Arithmetic operations
  - Addition
  - Subtraction
- Logic operatons
  - AND, OR, XOR ...
- Shift operations
  - Left, Right ...

Floating-point Numbers

$$X = \pm m \cdot b^e$$


```
b = 2
e = exponent - bias
bias = 2^(j-k-1)-1
if (m ≠ 0)then m = 1.mantissa
if (s = 0)then X > 0
if (s = 1)then X < 0
```

IEEE standard for floating-point arithmetic 754-2008. August 2008.

IEEE Specification

Precision	I	j	k	e(bits)	m(bits)	bias
Single	31	30	22	8	24	$127=2^7-1$
Double	63	62	51	11	53	$1023=2^{10}-1$

Single $(-1)^S \times (1.m) \times 2^{e-127}$

Double $(-1)^S \times (1.m) \times 2^{e-1023}$

IEEE standard for floating-point arithmetic 754-2008. August 2008.

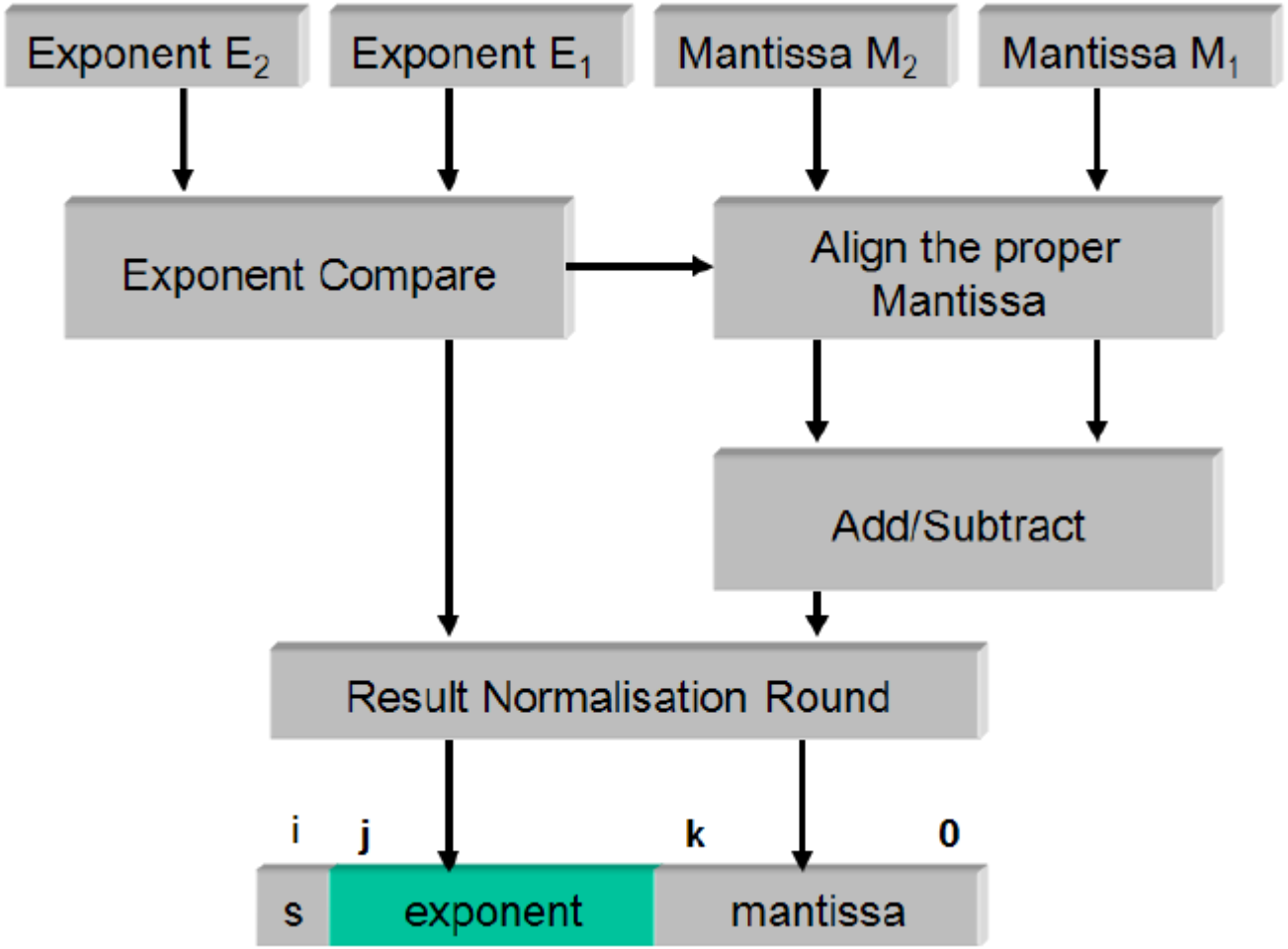
Exponent Compare & Align the Mantissa

$$X = 1.1100 \times 2^4 = 28$$
$$+Y = 1.1000 \times 2^2 = 6$$

$$1.1100 \times 2^4$$
Align $$+0.0110 \times 2^4$$
Add $$10.0010 \times 2^4 = 34$$

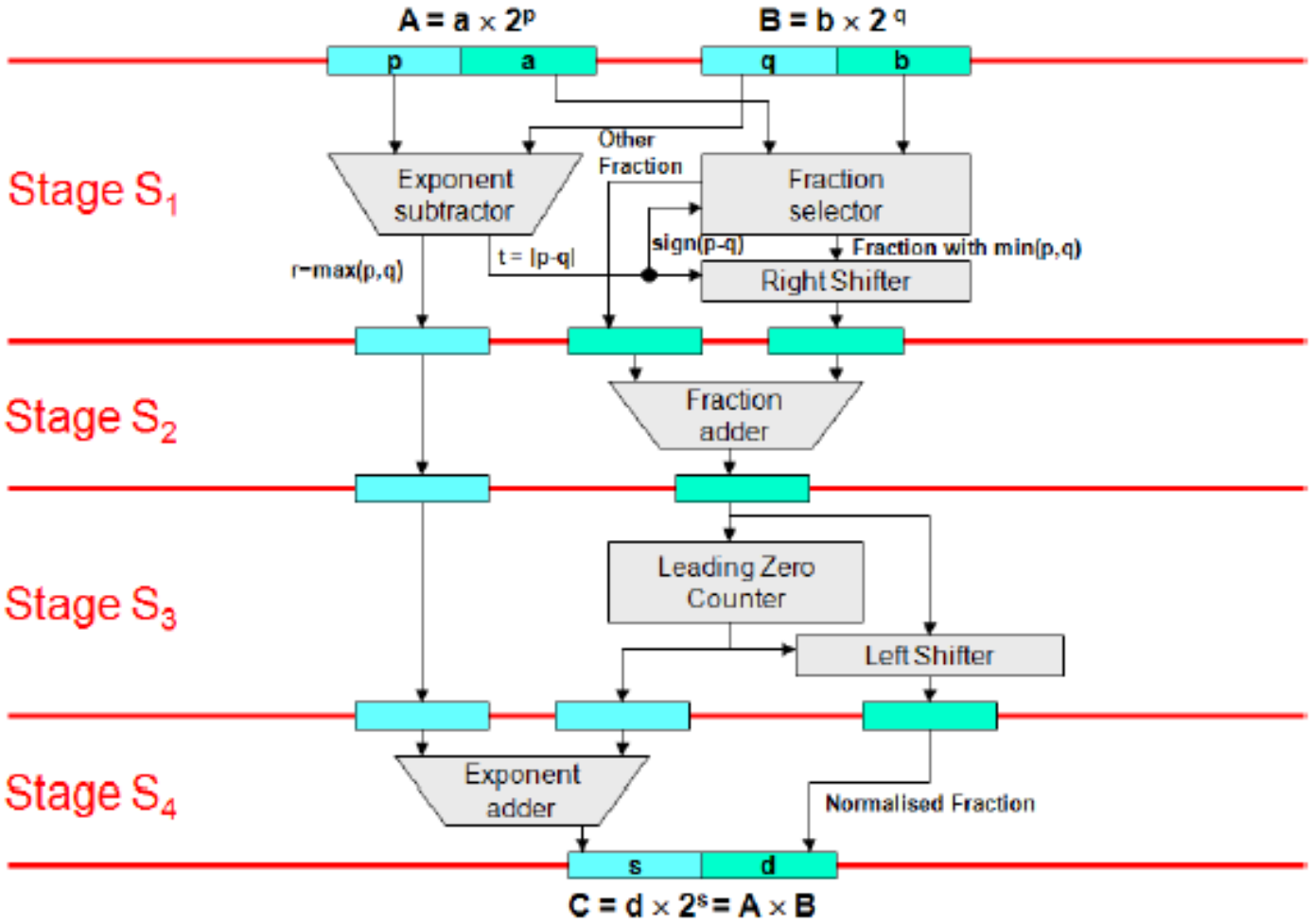
IEEE standard for floating-point arithmetic 754-2008. August 2008.

Floating-point Adder



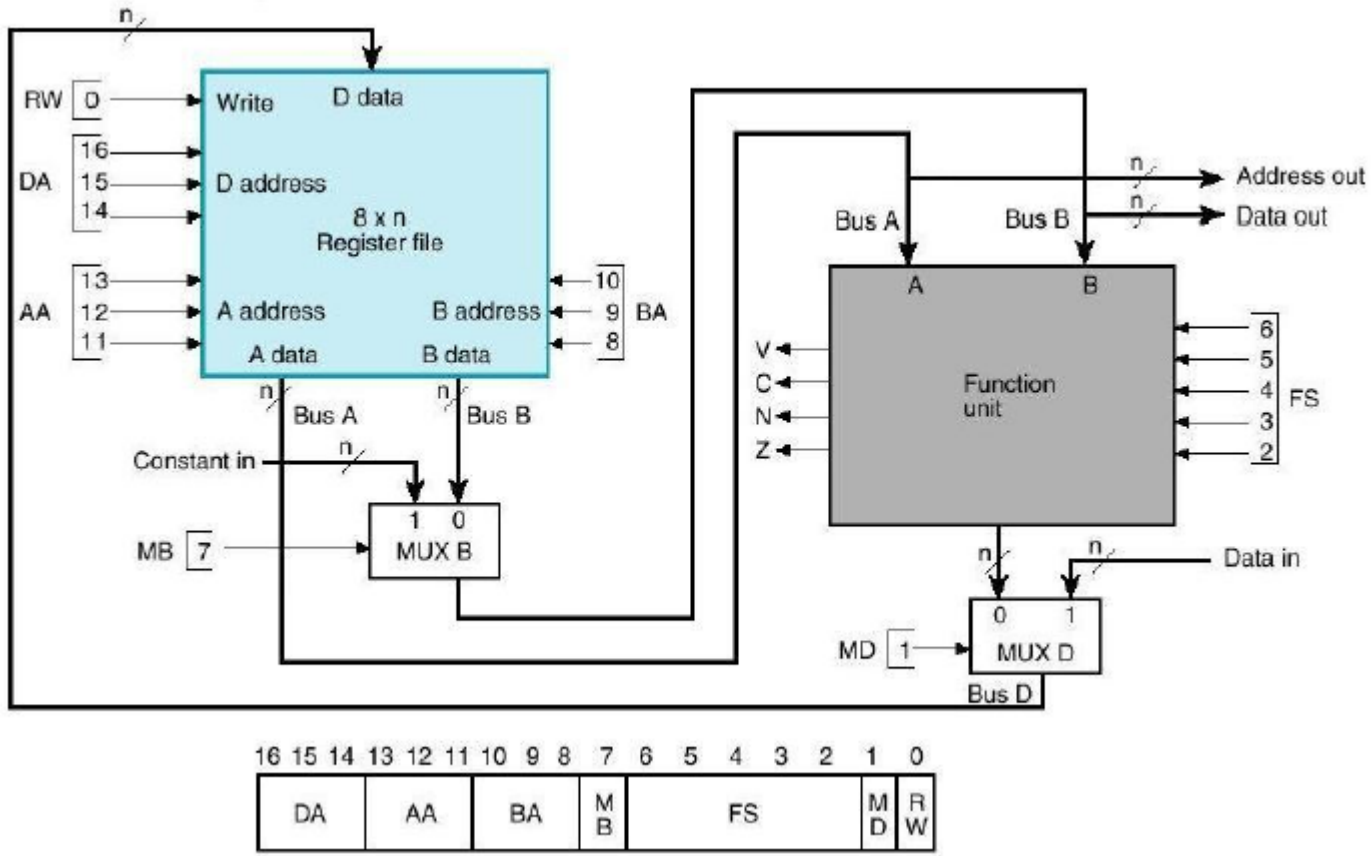
IEEE standard for floating-point arithmetic 754-2008. August 2008.

Pipelined Floating-point Adder



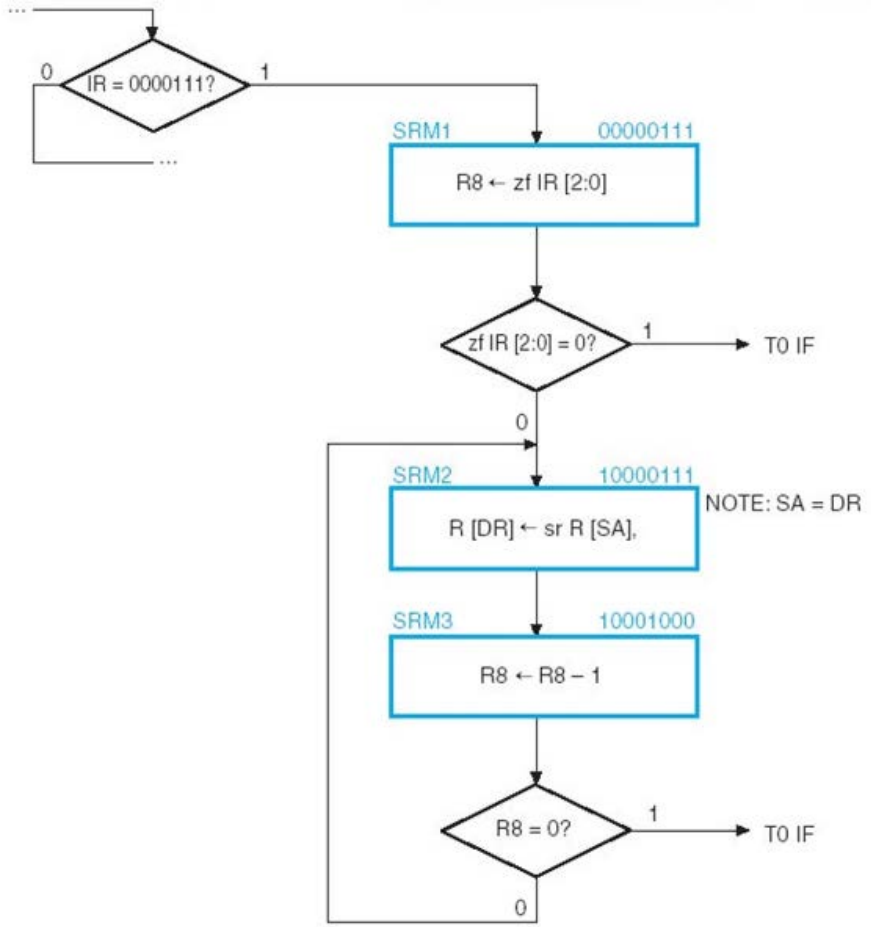
IEEE standard for floating-point arithmetic 754-2008. August 2008.

Datapath



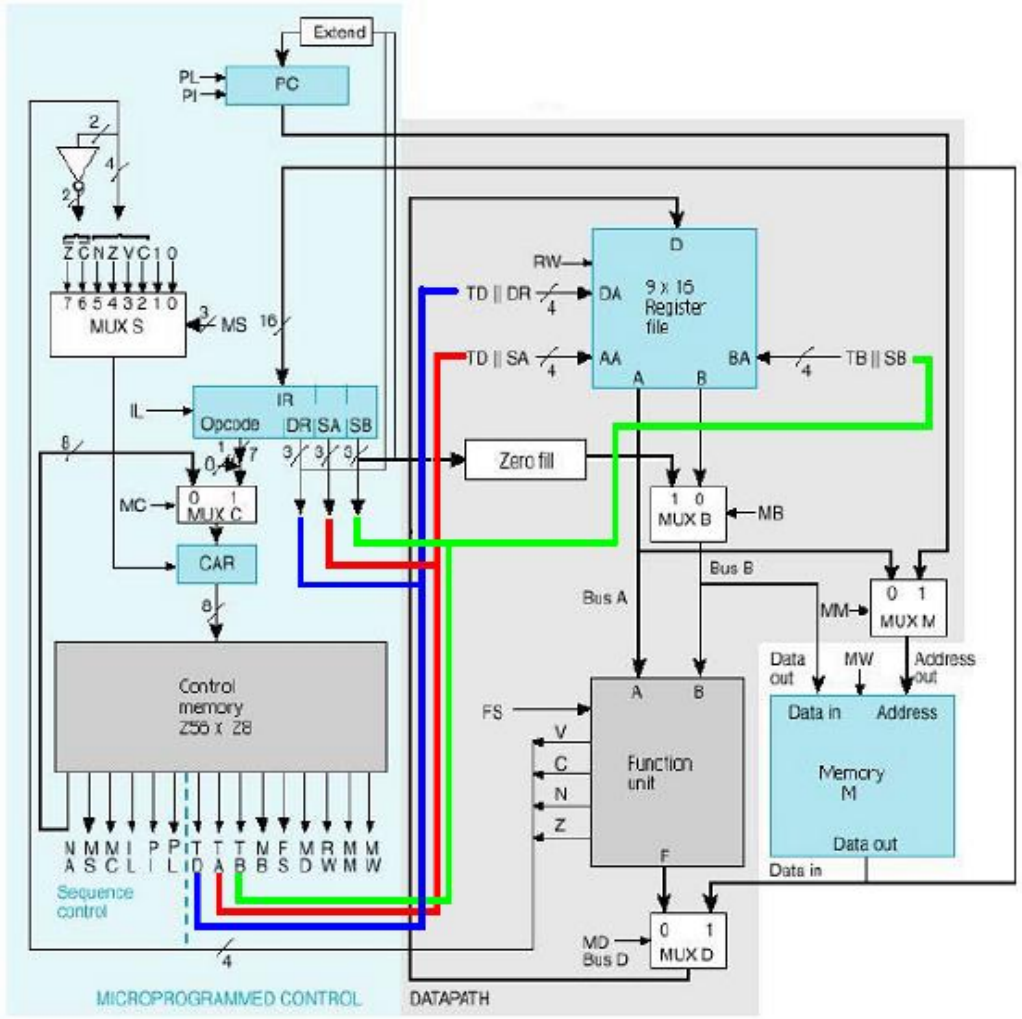
Morris Mano and Charles R. Kime. Logic and Computer Design Fundamentals. Prentice-Hall, second edition updated edition, 2001

Algorithmic State Machine



Morris Mano and Charles R. Kime. Logic and Computer Design Fundamentals. Prentice-Hall, second edition updated edition, 2001

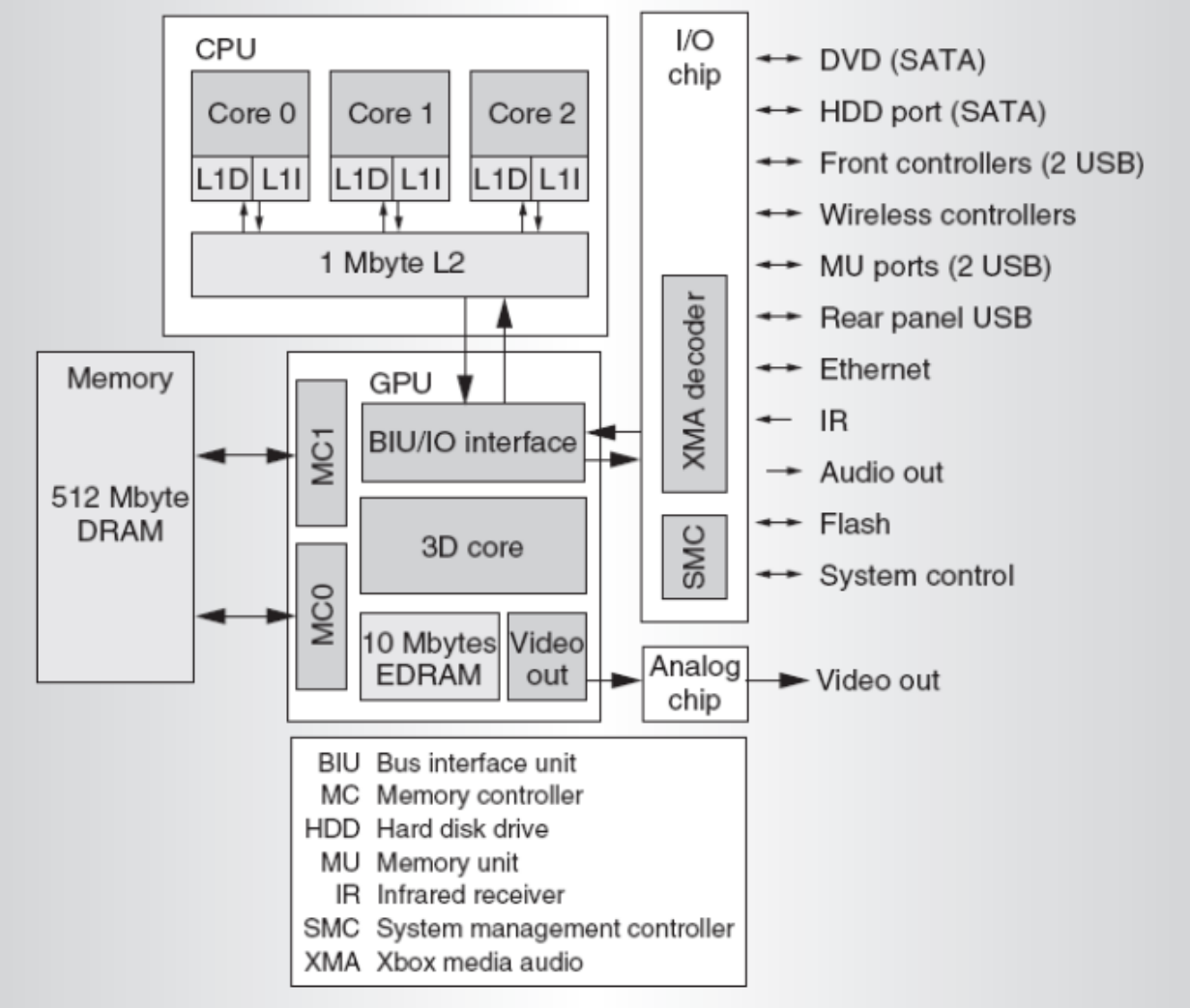
Control for the Datapath



Morris Mano and Charles R. Kime. Logic and Computer Design Fundamentals. Prentice-Hall, second edition updated edition, 2001

Xbox 360 (2005)

- CPU
  - 3 cores
- GPU
  - 48 parallel, unified shaders



Jeff Andrews and Nick Baker. Xbox 360 system architecture. IEEE Micro, 26(2):25–37, 2006

Amdahl's Law

- Amdahl's Law equation (1) allows us to compute the speedup that can be achieved.

$$\text{Speedup} = \frac{1}{\frac{\text{Fraction}_{\text{enhanced}}}{\text{Speedup}_{\text{enhanced}}} + \left(1 - \text{Fraction}_{\text{enhanced}}\right)} \quad (1)$$

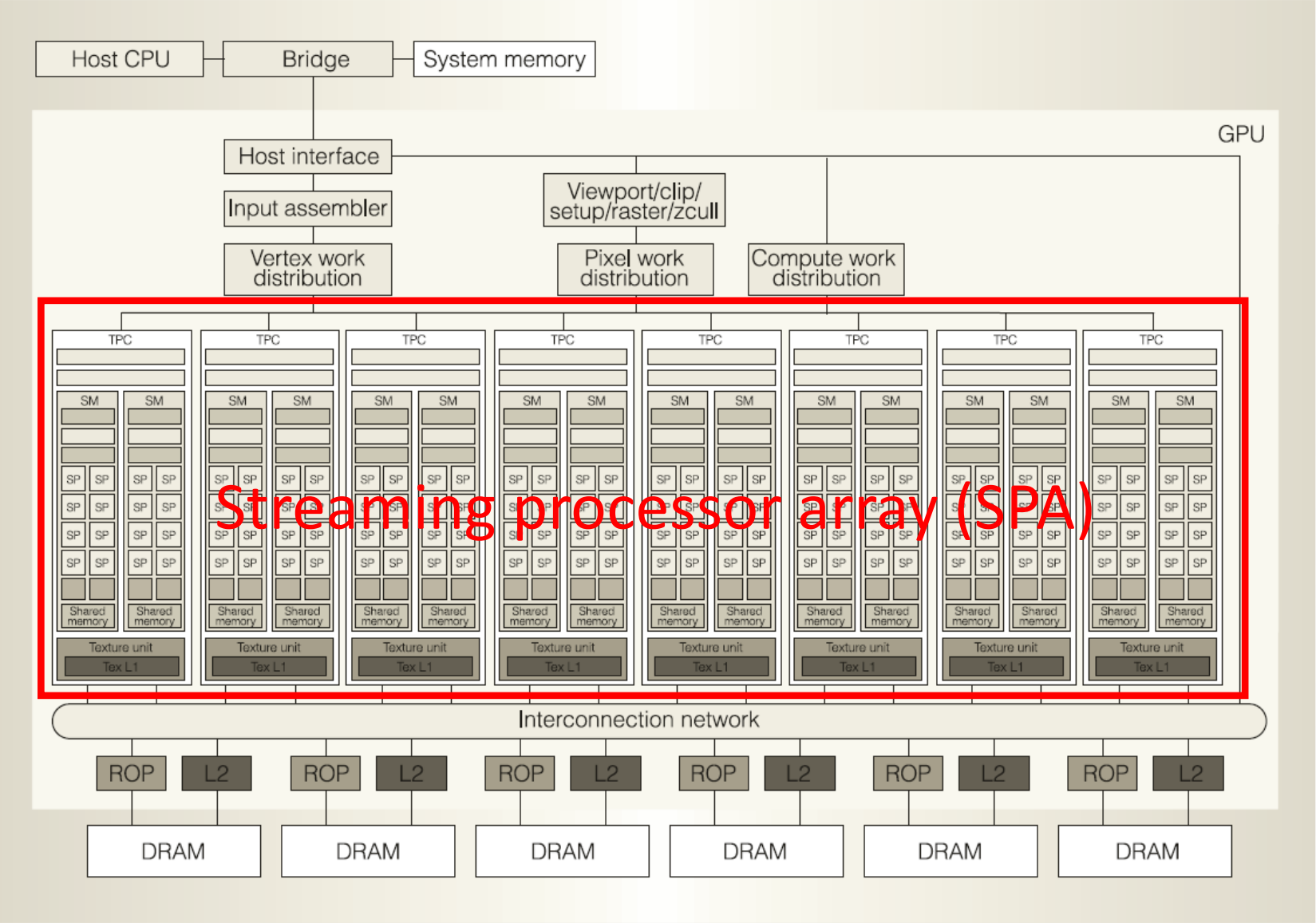
- Amdahl's Law equation (2) tells us that in order to achieve a speedup of 80 with 100 processors only 0.25% of the code may executed sequentially.

$$80 = \frac{1}{\frac{\text{Fraction}_{\text{parallel}}}{100} + \left(1 - \text{Fraction}_{\text{parallel}}\right)} \quad (2)$$

Gene M. Amdahl. 1967. Validity of the single processor approach to achieving large scale computing capabilities. In Proceedings of the April 18-20, 1967, spring joint computer conference (AFIPS '67 (Spring)). ACM, New York, NY, USA, 483-485. DOI=http://dx.doi.org/10.1145/1465482.1465560

NVIDIA Tesla (2008): A unified Graphics and Computing Architecture

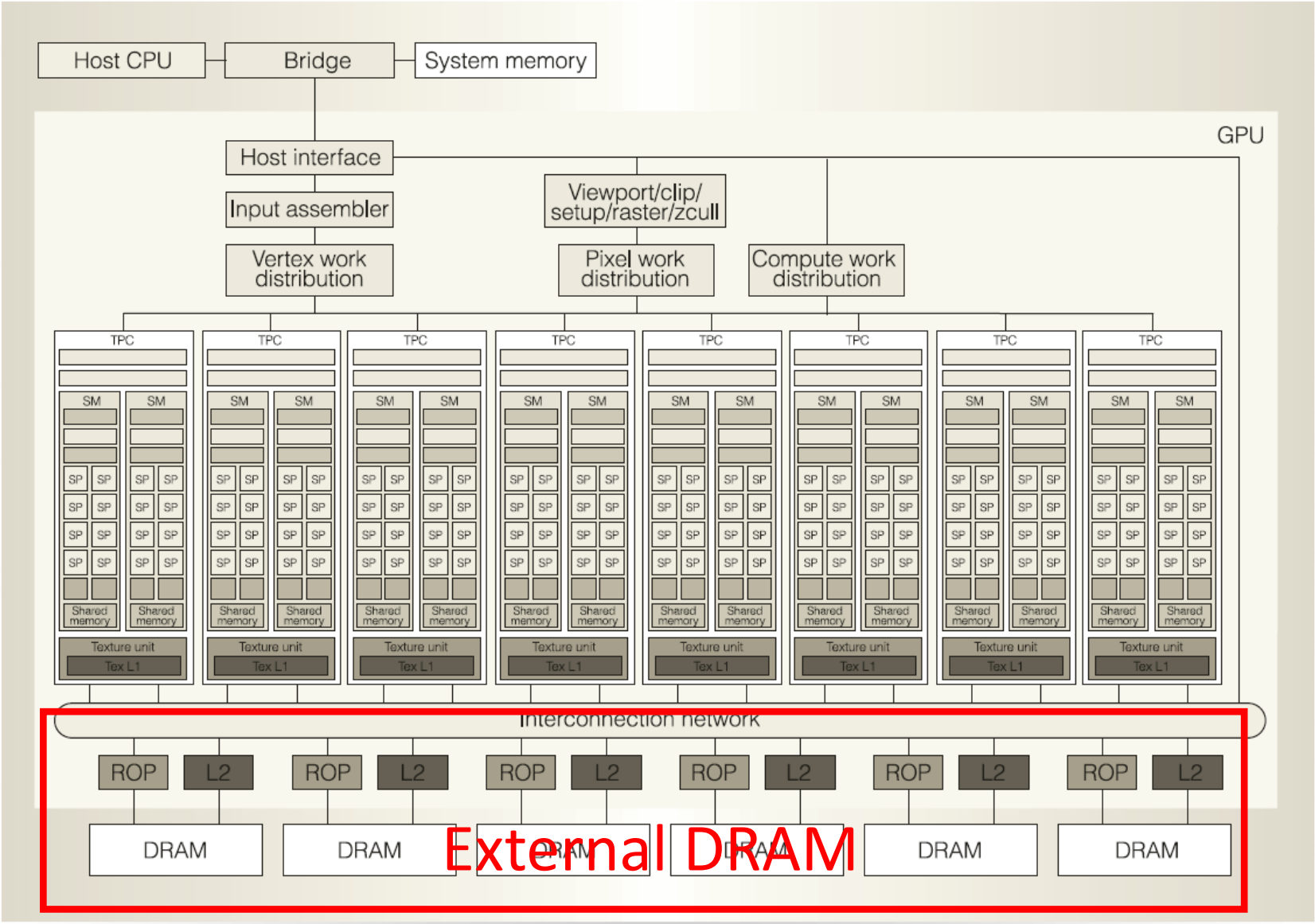
- Streaming-processors (SPs)
  - 128 (5120 in 2018)
- Streaming-multiprocessors (SMs)
  - 16
- Texture Processor Clusters (TPCs)
  - 8
- No logical order of graphics pipeline stages



E. Lindholm, J. Nickolls, S. Oberman and J. Montrym, "NVIDIA Tesla: A Unified Graphics and Computing Architecture," in IEEE Micro, vol. 28, no. 2, pp. 39-55, March-April 2008. doi: 10.1109/MM.2008.31

Scalable Memory System

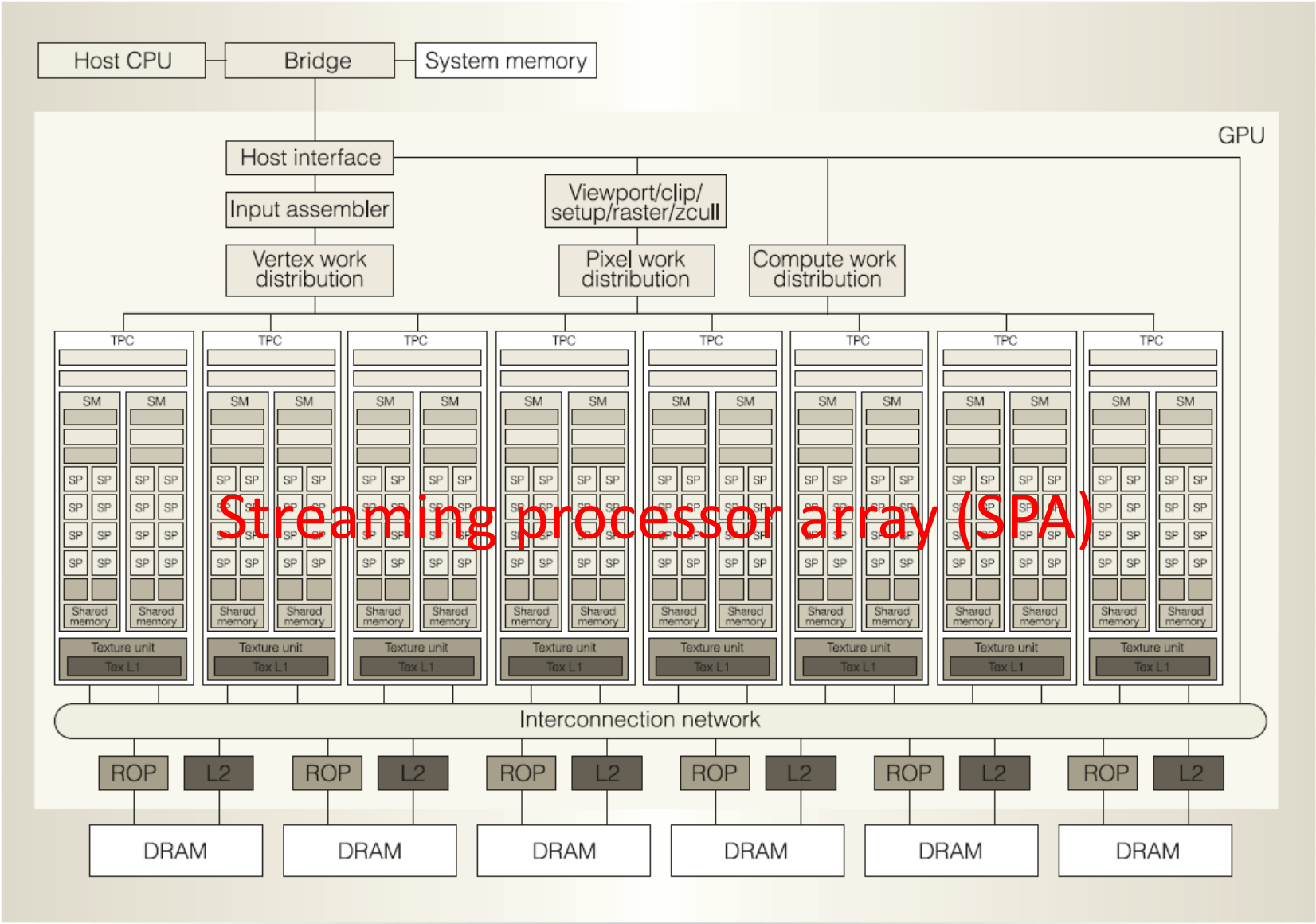
- Raster operation-processors (ROPs)
  - 6 (fixed-function)
  - Perform color and depth frame buffer directly on memory
- Interconnection Network
  - Moves fragment colour and depth from Streaming processor array (SPA) to Raster operation-processors (ROPs)
  - Also provides access to the textures in the external DRAM



E. Lindholm, J. Nickolls, S. Oberman and J. Montrym, "NVIDIA Tesla: A Unified Graphics and Computing Architecture," in IEEE Micro, vol. 28, no. 2, pp. 39-55, March-April 2008. doi: 10.1109/MM.2008.31

Input to the Streaming processor array (SPA)

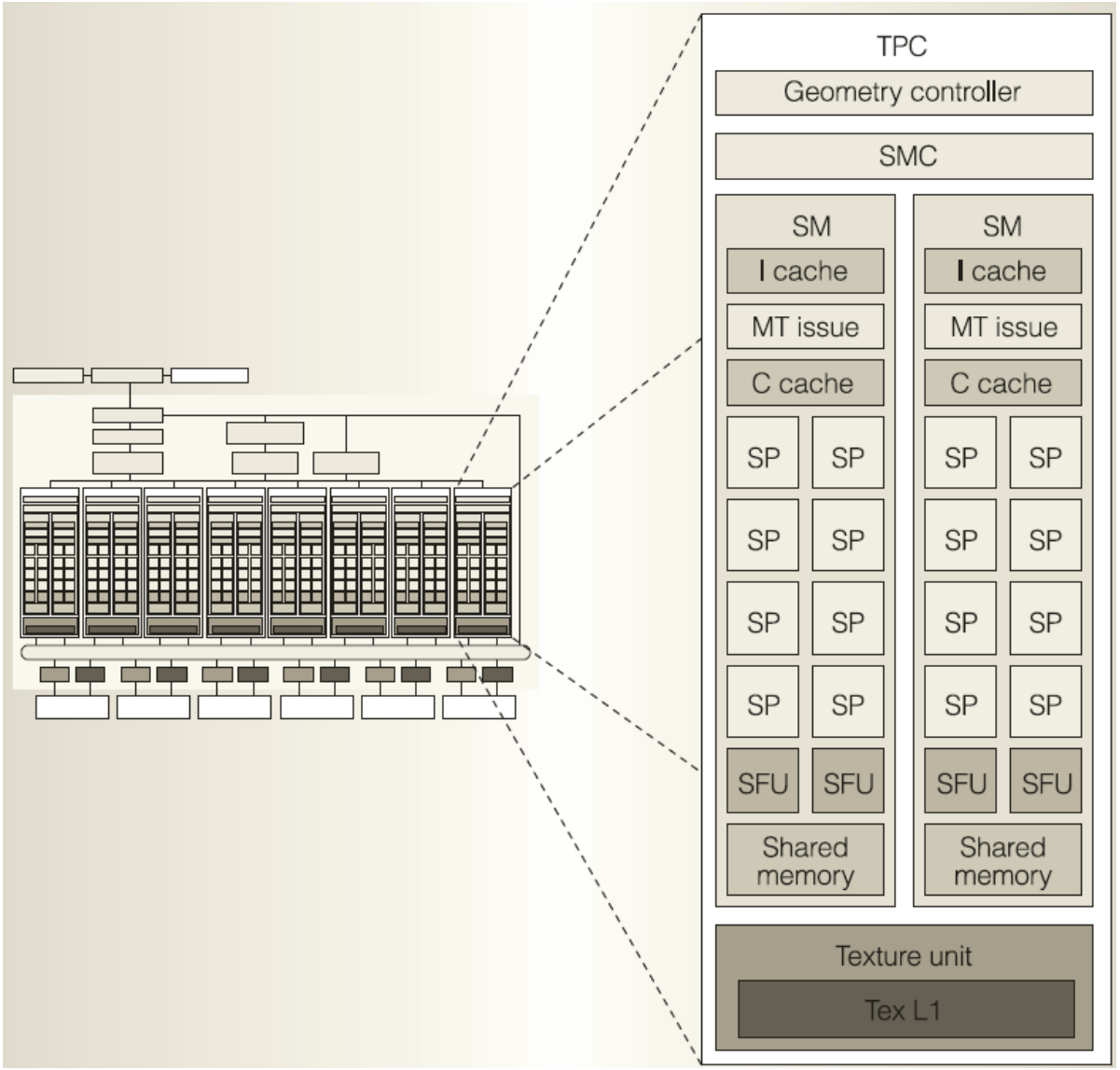
- Input assembler
  - Collects vertex work as directed by the input command stream
- Vertex work distribution
  - Distributes Vertex work packets to TPCs in the SPA
- The Texture Processor Clusters (TPCs) executes:
  - Vertex shader programs
  - Geometry shader programs
  - Output is written to on-chip buffers
- Viewport/clip/setup/raster/Zcull
  - Takes the output from the buffer
  - Rasterizes it into fragments
- Pixel work distribution
  - Send fragments to the appropriate TCDs



E. Lindholm, J. Nickolls, S. Oberman and J. Montrym, "NVIDIA Tesla: A Unified Graphics and Computing Architecture," in IEEE Micro, vol. 28, no. 2, pp. 39-55, March-April 2008. doi: 10.1109/MM.2008.31

Texture Processor Cluster (TPC)

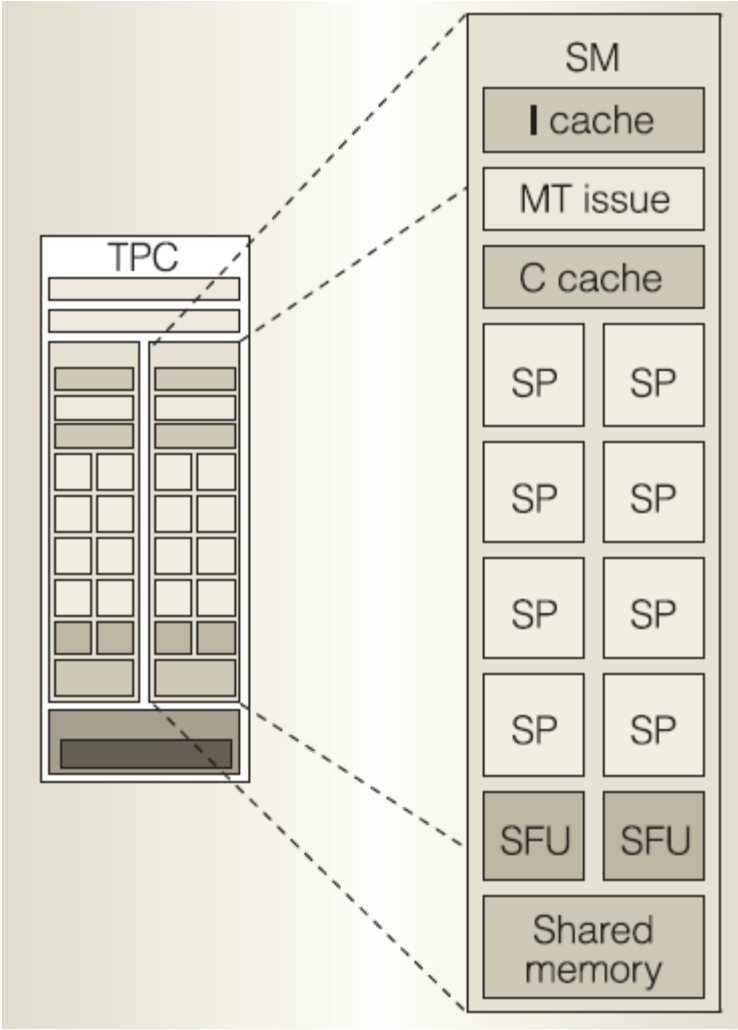
- Geometry Controller
  - Manages input and output vertex attributes
- SM Controller (SMC)
- Streaming multiprocessors (SMs)
  - Two
- Texture Unit



E. Lindholm, J. Nickolls, S. Oberman and J. Montrym, "NVIDIA Tesla: A Unified Graphics and Computing Architecture," in IEEE Micro, vol. 28, no. 2, pp. 39-55, March-April 2008. doi: 10.1109/MM.2008.31

Streaming Multiprocessor (SM)

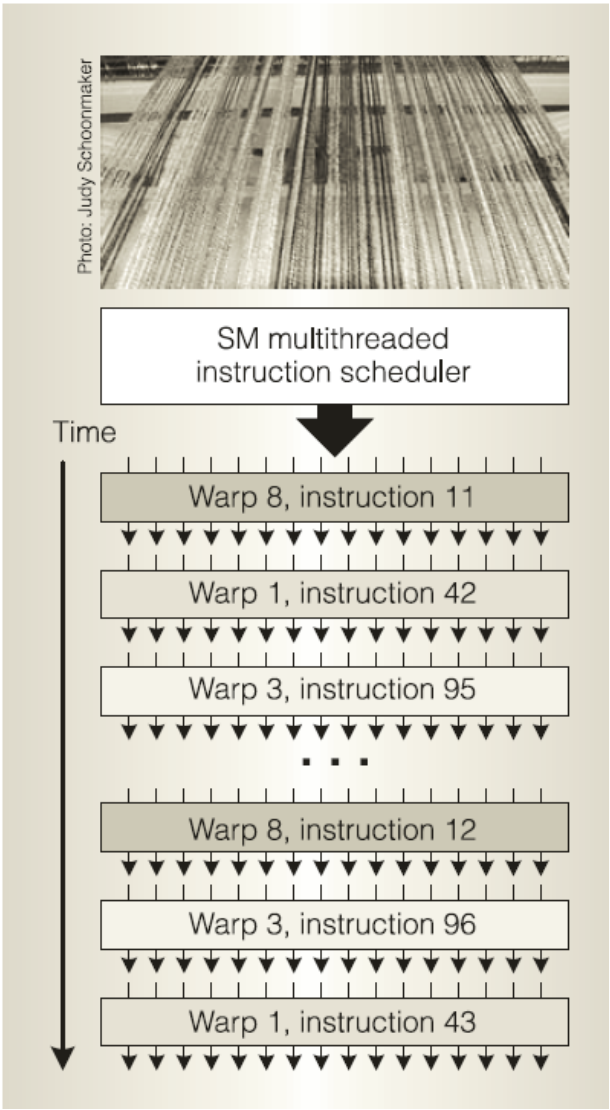
- Unified graphics and computing multiprocesoor
  - Vertex shader
  - Geometry shader
  - Fragment shader
  - Parallel computing shader
- Special function Units (SFUs)
- Multithreaded instruction fetch and issue unit (MT Issue)



E. Lindholm, J. Nickolls, S. Oberman and J. Montrym, "NVIDIA Tesla: A Unified Graphics and Computing Architecture," in IEEE Micro, vol. 28, no. 2, pp. 39-55, March-April 2008. doi: 10.1109/MM.2008.31

Single-instruction Multiple-thread

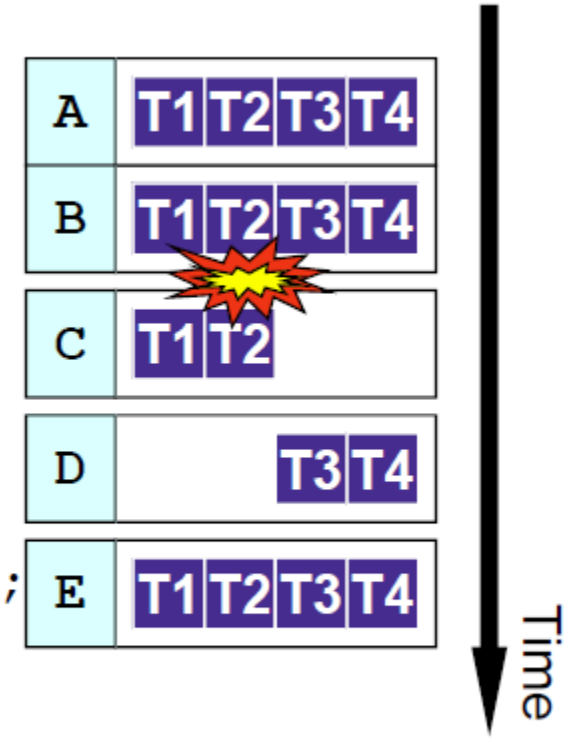
- Warps
  - 32 parallel threads
  - Pool of 24 warps
  - Total of 768



E. Lindholm, J. Nickolls, S. Oberman and J. Montrym, "NVIDIA Tesla: A Unified Graphics and Computing Architecture," in IEEE Micro, vol. 28, no. 2, pp. 39-55, March-April 2008. doi: 10.1109/MM.2008.31

SIMT Stack

```
foo[] = {4,8,12,16};
 A: v = foo[tid.x];
 B: if (v < 10)
 C:     v = 0;
    else
 D:     v = 10;
 E: w = bar[tid.x]+v;
```



One stack per warp

SIMT Stack

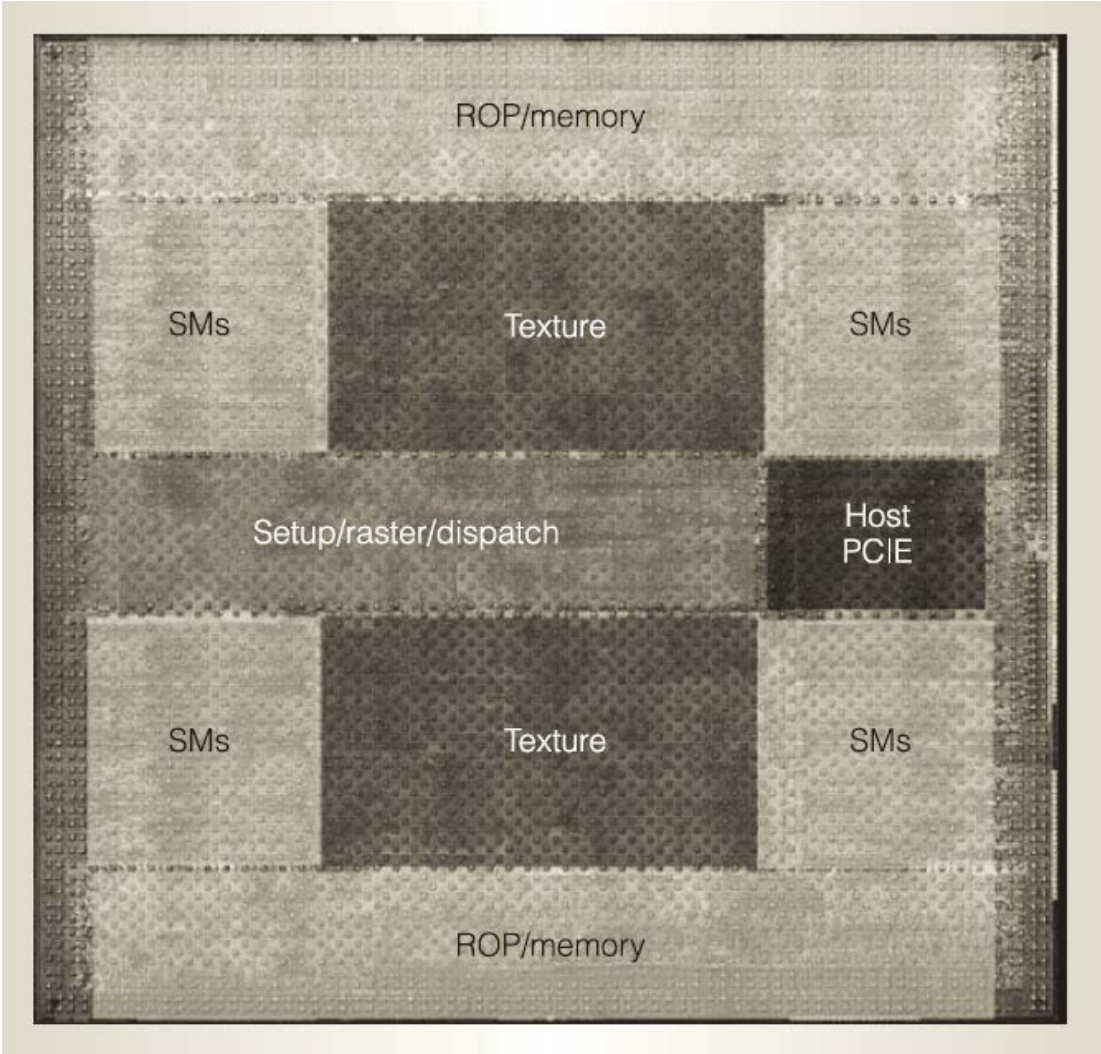
PC	RPC	Active Mask
E	-	1111
D	E	0011
C	E	1100

Handles *Branch Divergence*

GPGPU-Sim 3.x: A Performance Simulator for Manycore Accelerator Research, 2012 http://www.gpgpu-sim.org/micro2012-tutorial/

GeForce 8800 Ultra Die Layout

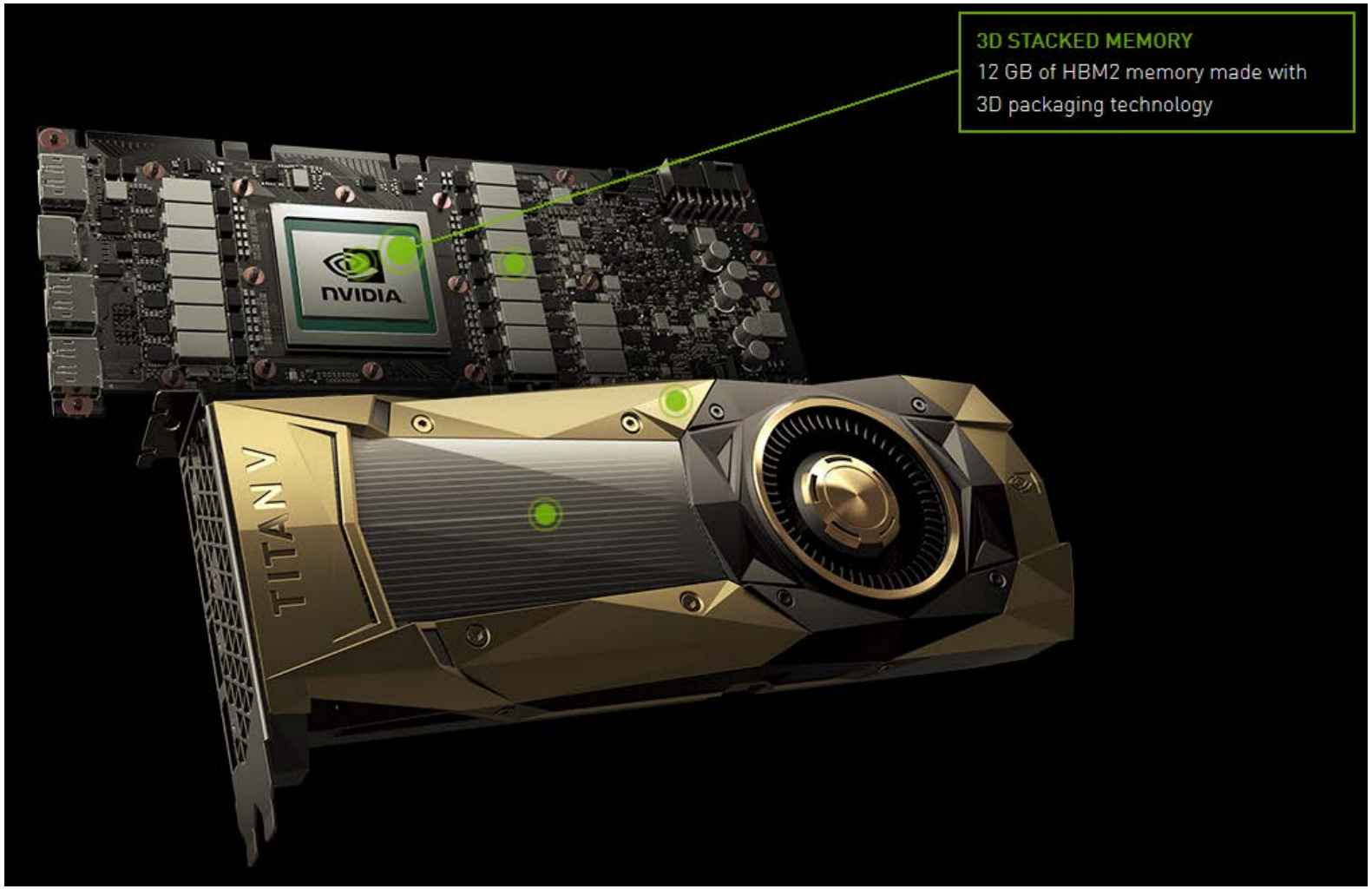
- Text



E. Lindholm, J. Nickolls, S. Oberman and J. Montrym, "NVIDIA Tesla: A Unified Graphics and Computing Architecture," in IEEE Micro, vol. 28, no. 2, pp. 39-55, March-April 2008. doi: 10.1109/MM.2008.31

NVIDIA TITAN (2018)

- Transistor Count
  - 21.1 Billion
- Total Video Memory
  - 12288 MB HBM2
- CUDA Cores
  - 5120 (single precision)
- 640 Tensor Cores
- Streaming Multiprocessors (SMs)
  - 80 (64 SP per SM)
- 110 TeraFLOPS



https://www.nvidia.com/en-us/titan/titan-v/