Muti- and Many-Core Discussion

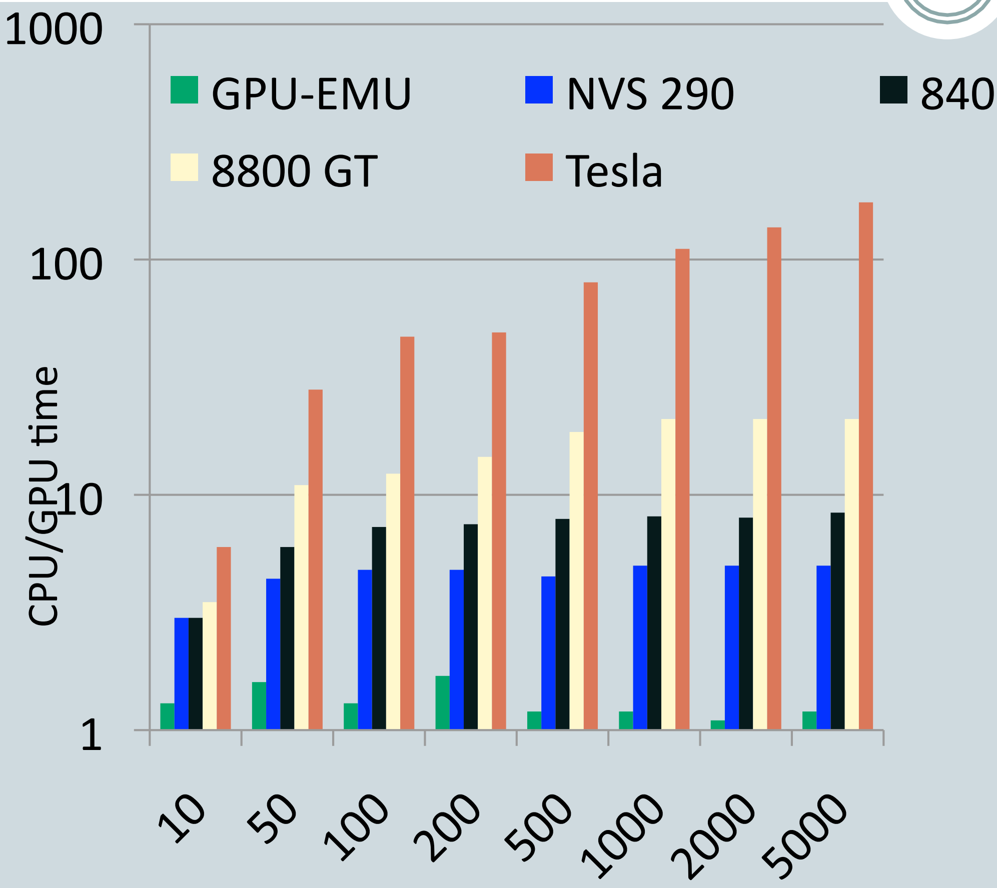
MOHAMMAD AL-TURANY
GSI DARMSTADT

Software for Multi- and Many-core

- We have to produce software that transparently scale its parallelism to balance the increasing number of CPU/GPU cores

- 3D graphics games transparently scale their parallelism to almost any number of GPU cores without problems! Why not in physics software?

Runge-Kutta: Gain for different cards



Trk/ Event	GPU emu	NVS 290 (16)	8400 GT (32)	8800 GT (112)	Tesla (240)
10	1.30	3	3	3.5	6
50	1.60	4.4	6	11	28
100	1.30	4.8	7.3	12.3	47
200	1.70	4.8	7.5	14.5	49
500	1.20	4.5	7.9	18.5	80
1000	1.20	5	8.1	21	111
2000	1.10	5	8	21	137
5000	1.20	5	8.4	21	175

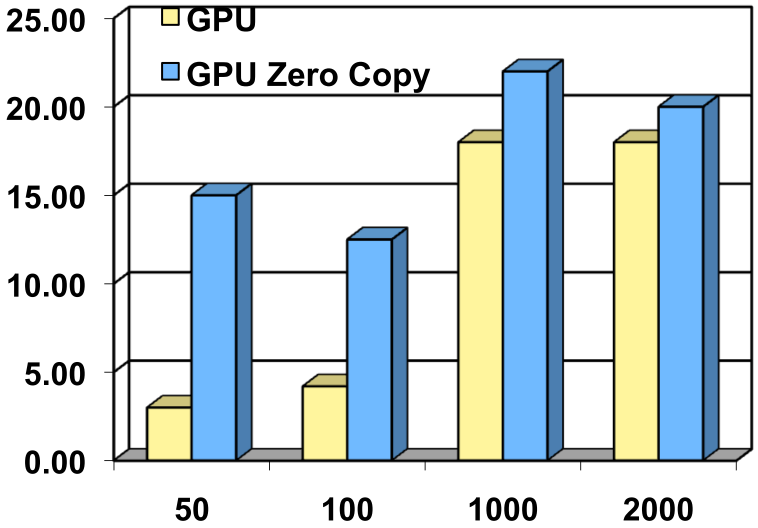
DETAILS: FRIDAY, 26.02

Applying CUDA computing model to event reconstruction software

Track + vertex fitting on CPU and GPU

CPU Time/GPU Time

Track/Event	50	100	1000	2000
GPU	3.0	4.2	18	18
GPU (Zero Copy)	15	13	22	20



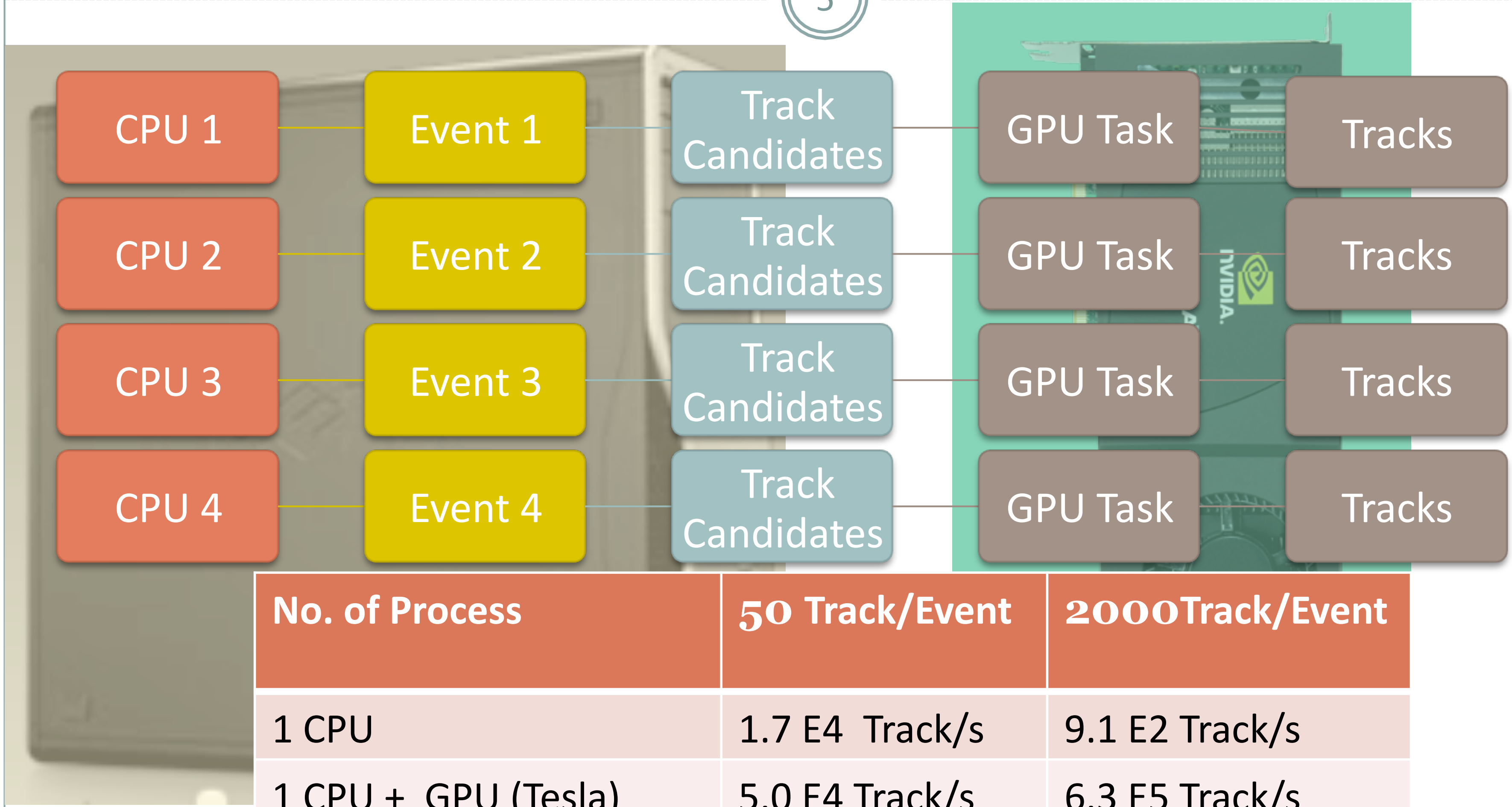
Time needed per event (ms)

	50	100	1000	2000
CPU	3.0	5.0	120	220
GPU	1.0	1.2	6.5	12.5
GPU (Zero Copy)	0.2	0.4	5.4	10.5

DETAILS: FRIDAY, 26.02

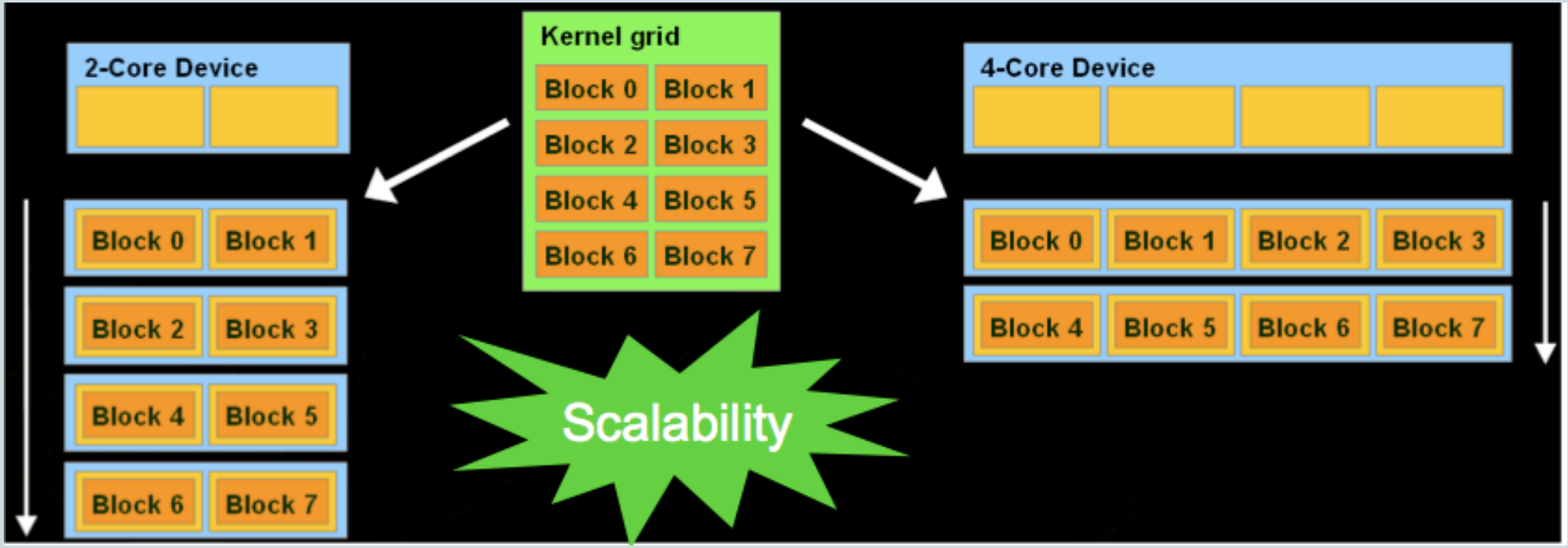
Applying CUDA computing model to event reconstruction software

Parallelization on CPU/GPU

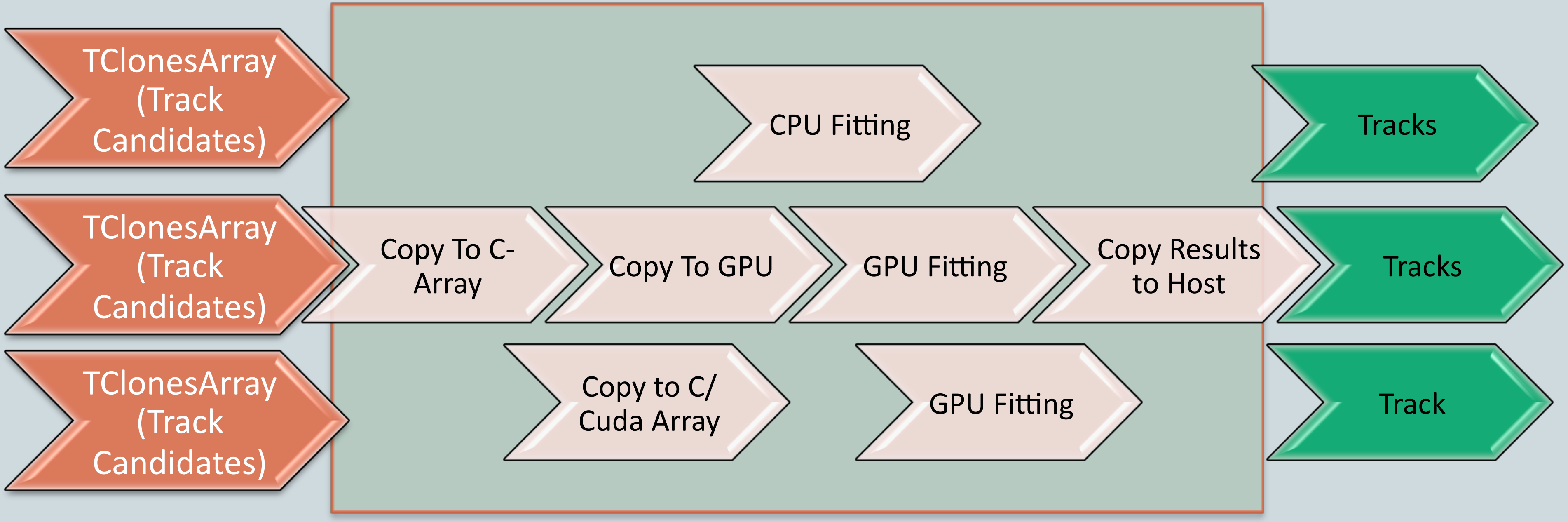


No. of Process	50 Track/Event	2000Track/Event
1 CPU	1.7 E4 Track/s	9.1 E2 Track/s
1 CPU + GPU (Tesla)	5.0 E4 Track/s	6.3 E5 Track/s
4 CPU + GPU (Tesla)	1.2 E5 Track/s	2.2 E6 Track/s

Scalability in CUDA



Comparisons between different techniques



Using the GPUs include some overhead in data processing which has to be considered in the comparisons to CPU code

NVIDIA's Next Generation CUDA Architecture

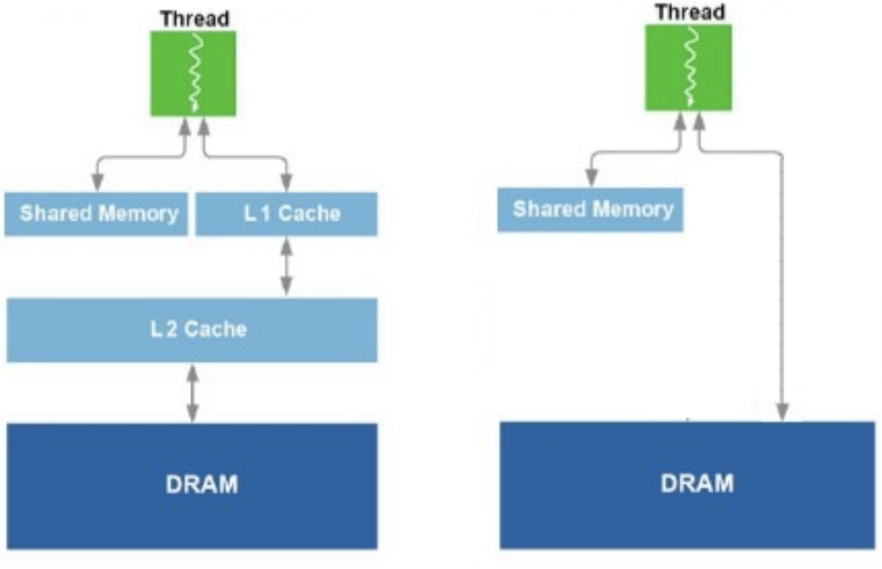
FERMI

Features:

Support a true cache hierarchy in combination with on-chip shared memory

Improves bandwidth and reduces latency through L1 cache's configurable shared memory

Fast, coherent data sharing across the GPU through unified L2 cache



Fermi

Tesla

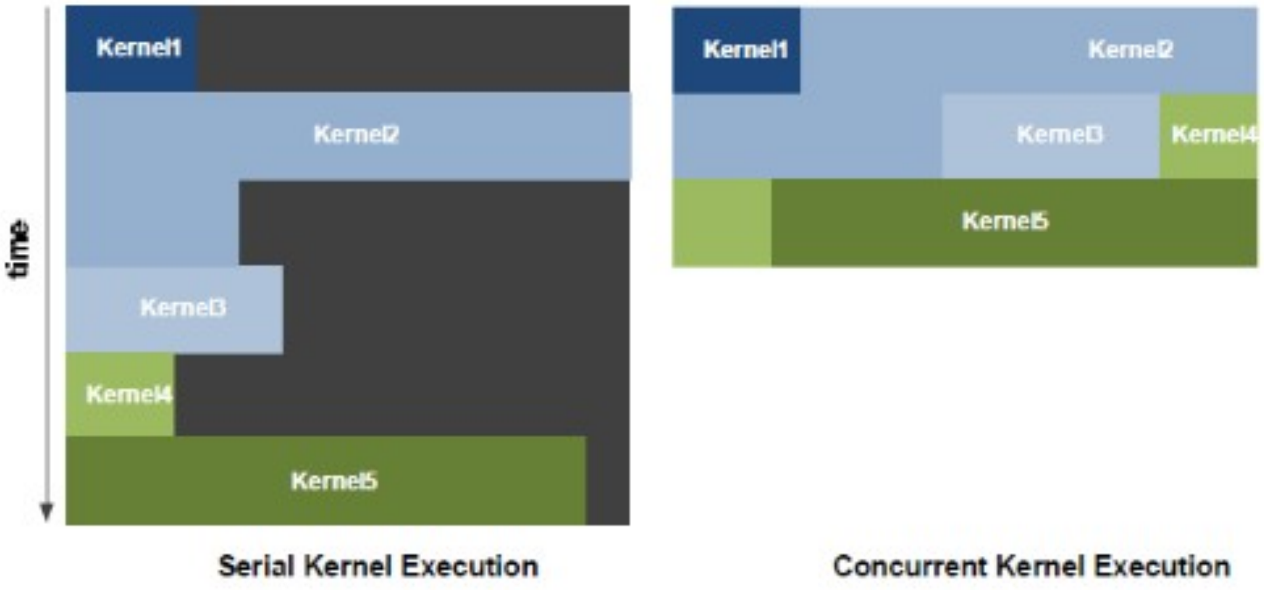
http://www.behardware.com/art/imprimer/772/

NVIDIA GigaThread™ Engine

Increased efficiency with concurrent kernel execution

Dedicated, bi-directional data transfer engines

Intelligently manage tens of thousands of threads



http://www.behardware.com/art/imprimer/772/

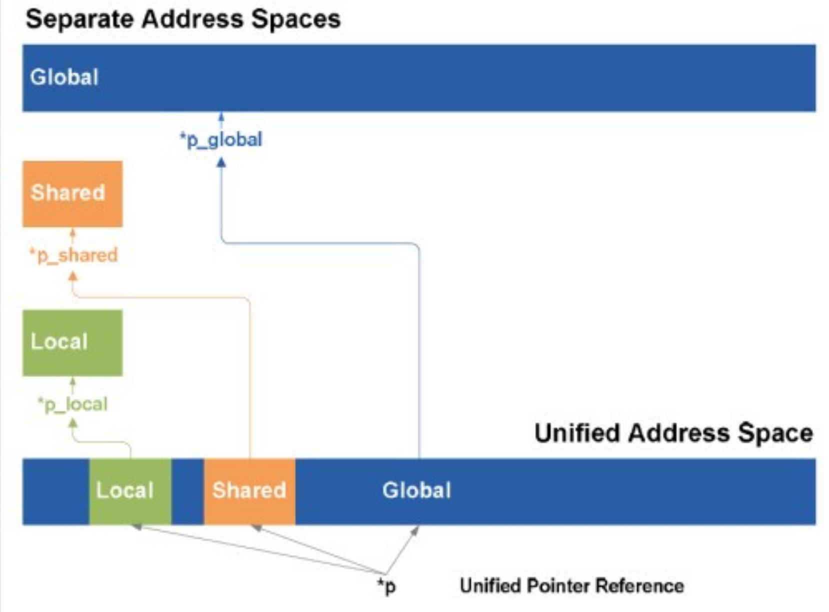
ECC Support

- First GPU architecture to support ECC
- Detects and corrects errors before system is affected
- Protects register files, shared memories, L1 and L2 cache, and DRAM

Unified address space

Groups local, shared and global memory in the same address space.

This unified address space means support for pointers and object references that are necessary for high-level languages such as C++.



http://www.behardware.com/art/imprimer/772/

Conclusion

- With Fermi we are getting towards the end of the distinction between CPUs and GPUs
  - ○ The GPU increasingly taking on the form of a massively parallel co-processor