RUPRECHT-KARLS-
UNIVERSITÄT
HEIDELBERG

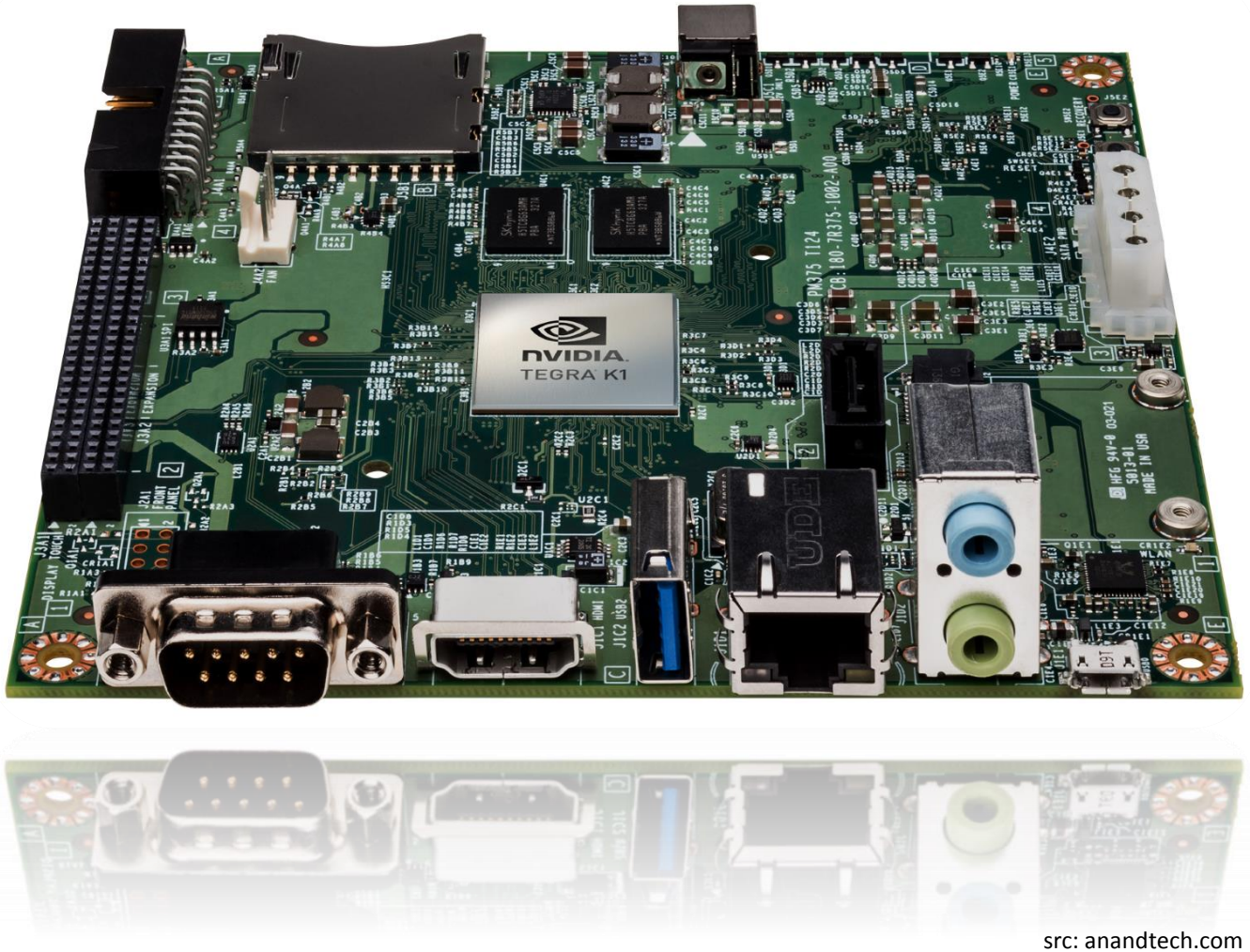
Jetson TK1

Seminararbeit

Benjamin Baumann

Contents

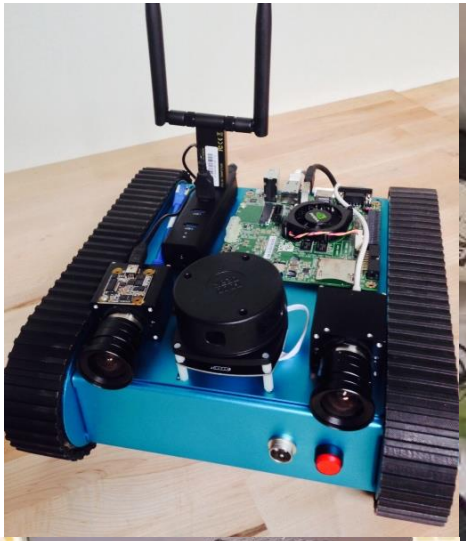
- Field of Application
- Jetson TK1
- GPU Basics
  - Architecture
  - CUDA
- Benchmark
  - Performance
  - Energy Efficiency
- Related Work
- Future
- Conclusion

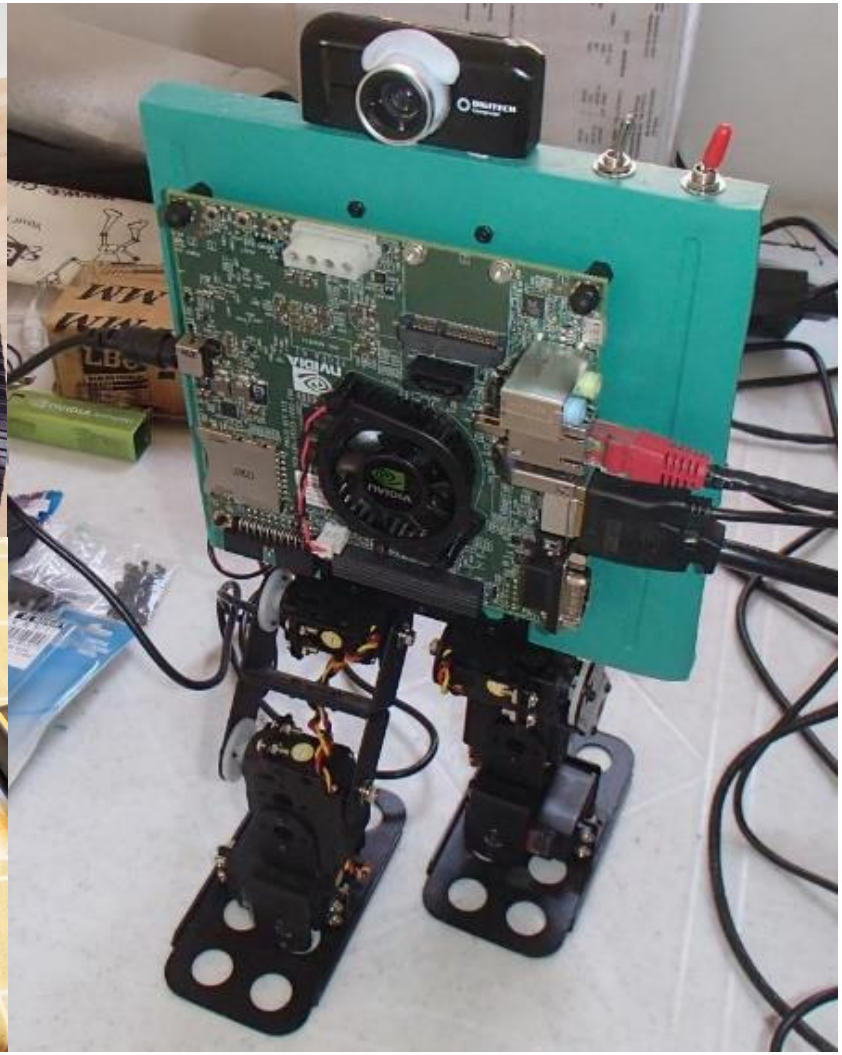


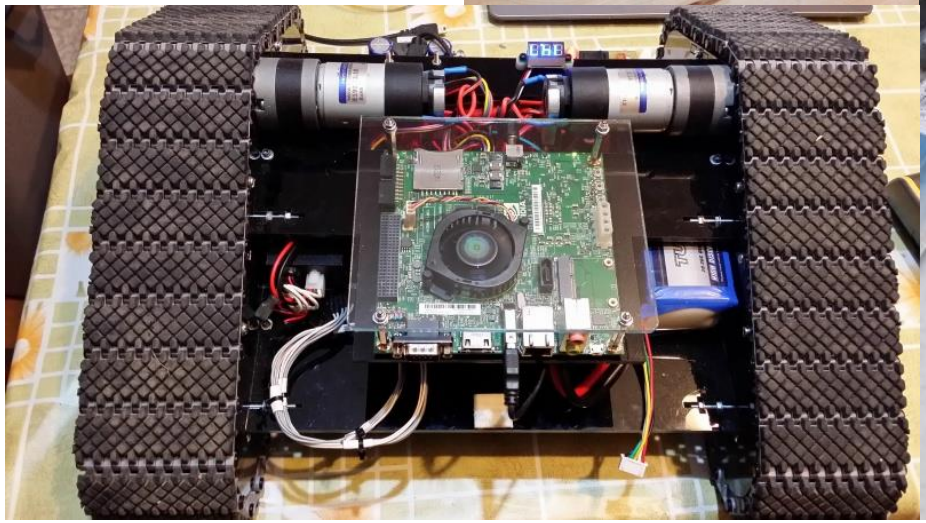
src: anandtech.com

Field of Application

- Robotics



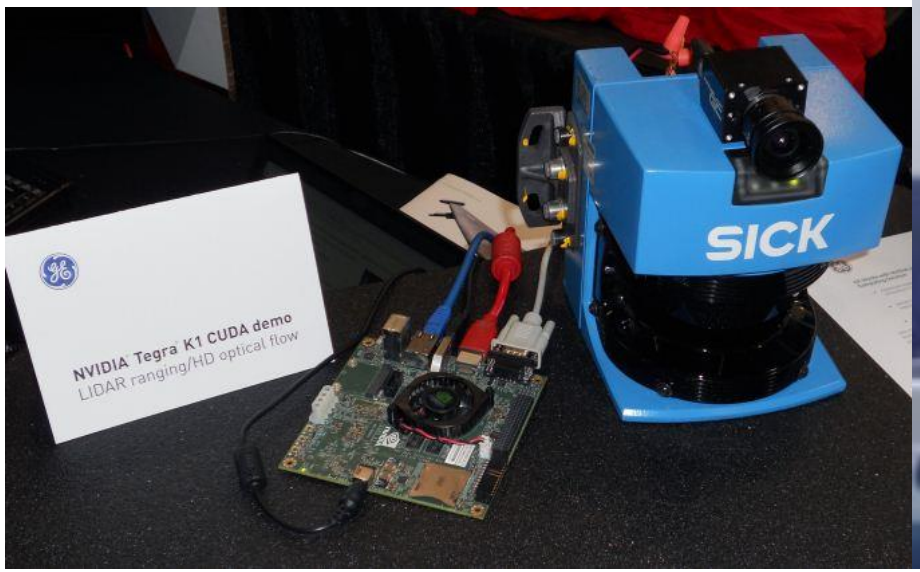


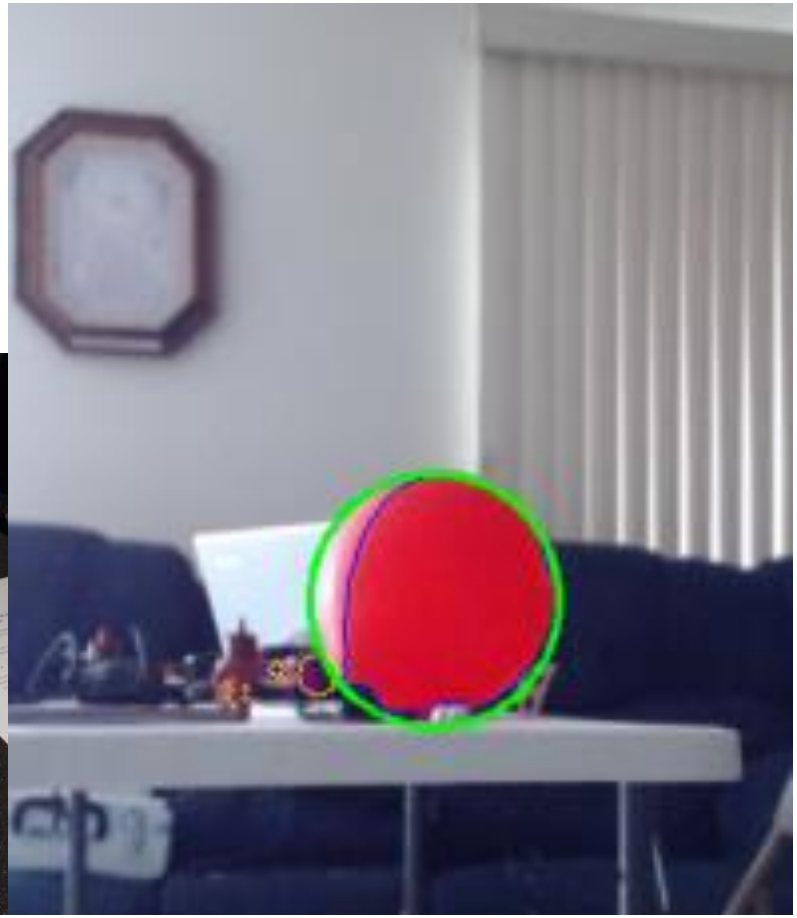


src: elinux.org

Field of Application

- Image Processing
  - Object detection
  - Computer Vision

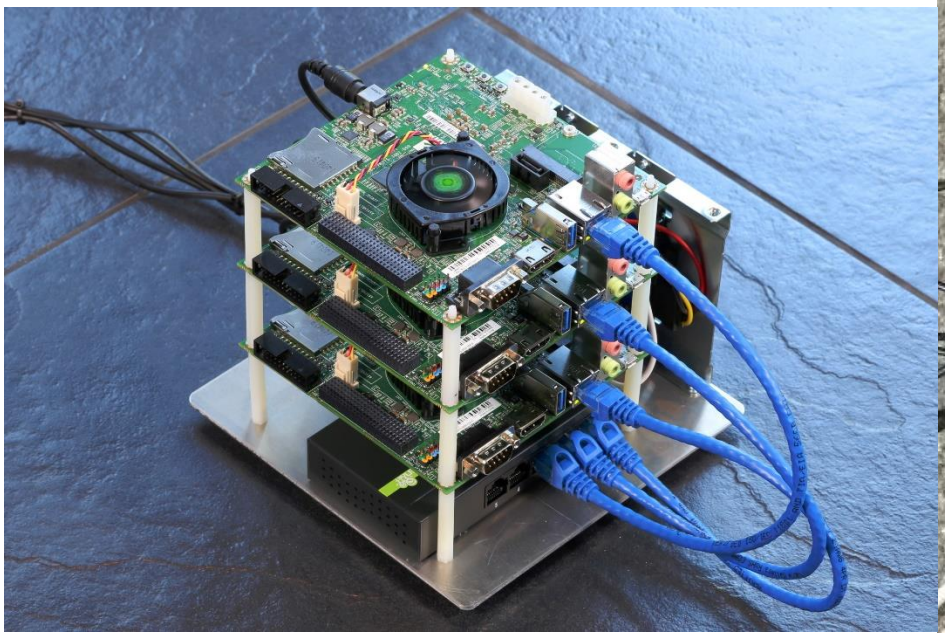




src: elinux.org

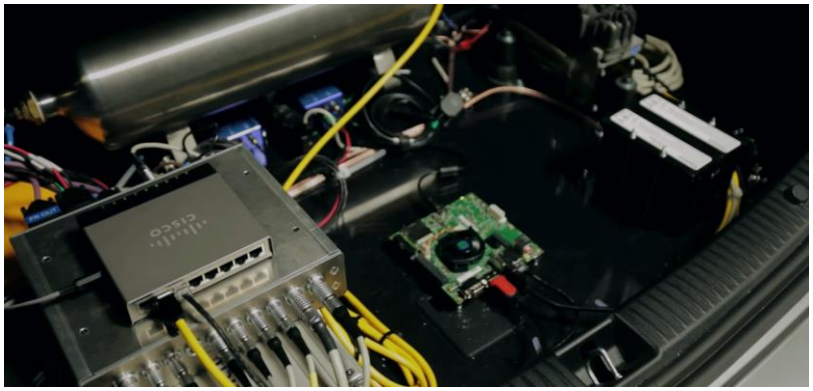
Field of Application

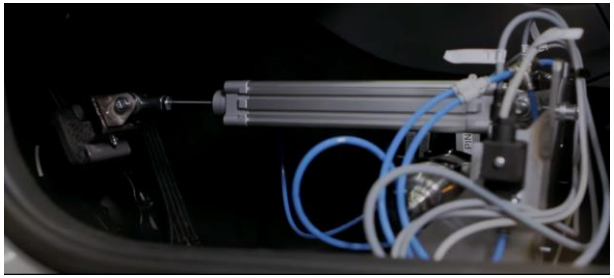
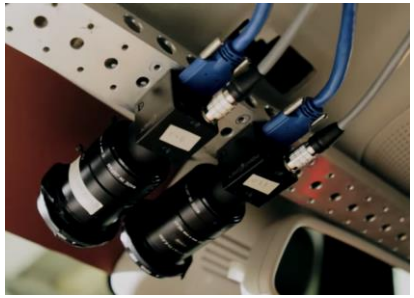
- Distributed computing



src: elinux.org

AdasWorks Automated Driving



- Automated Driving using a Jetson TK1

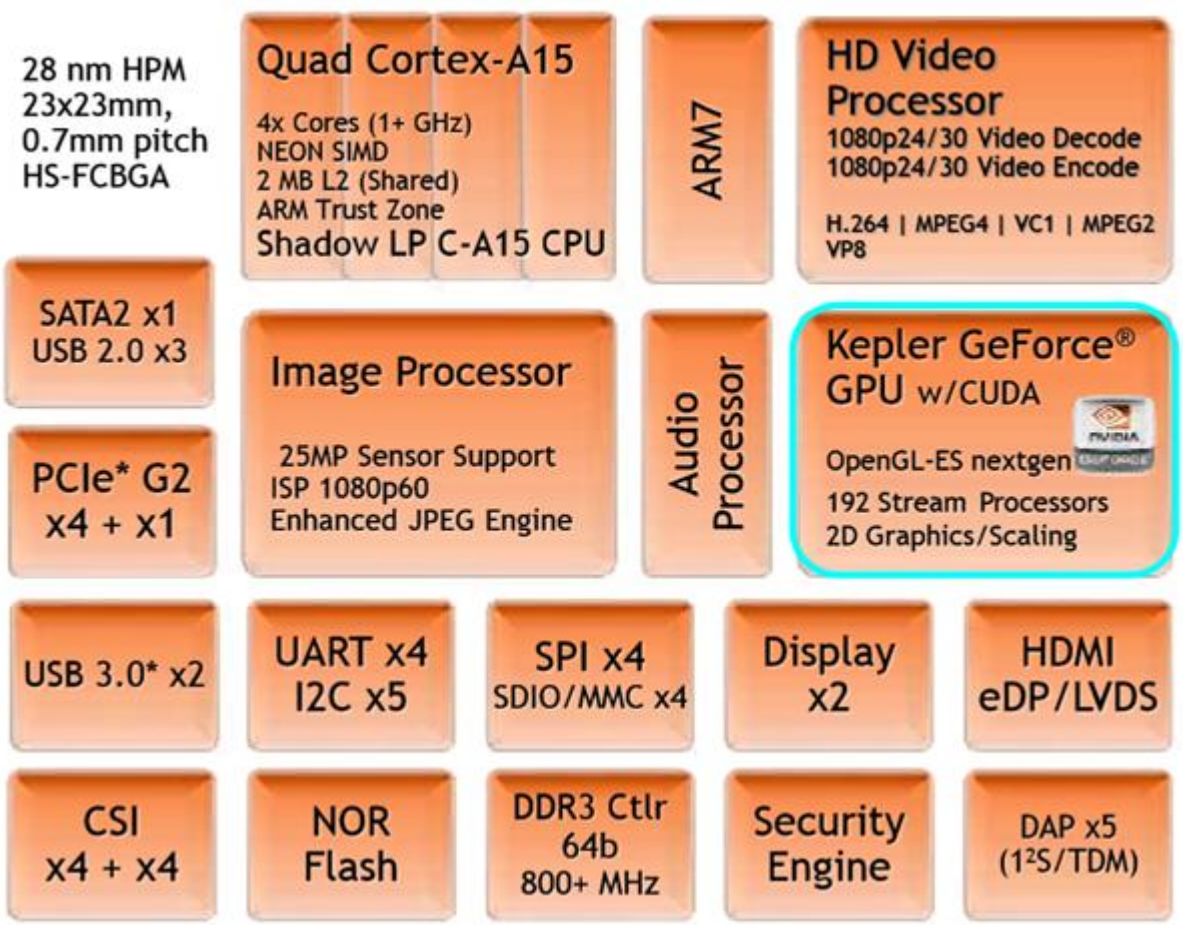
src: www.youtube.com/watch?v=37cOQS9gc1w

Jetson TK1

src: anandtech.com

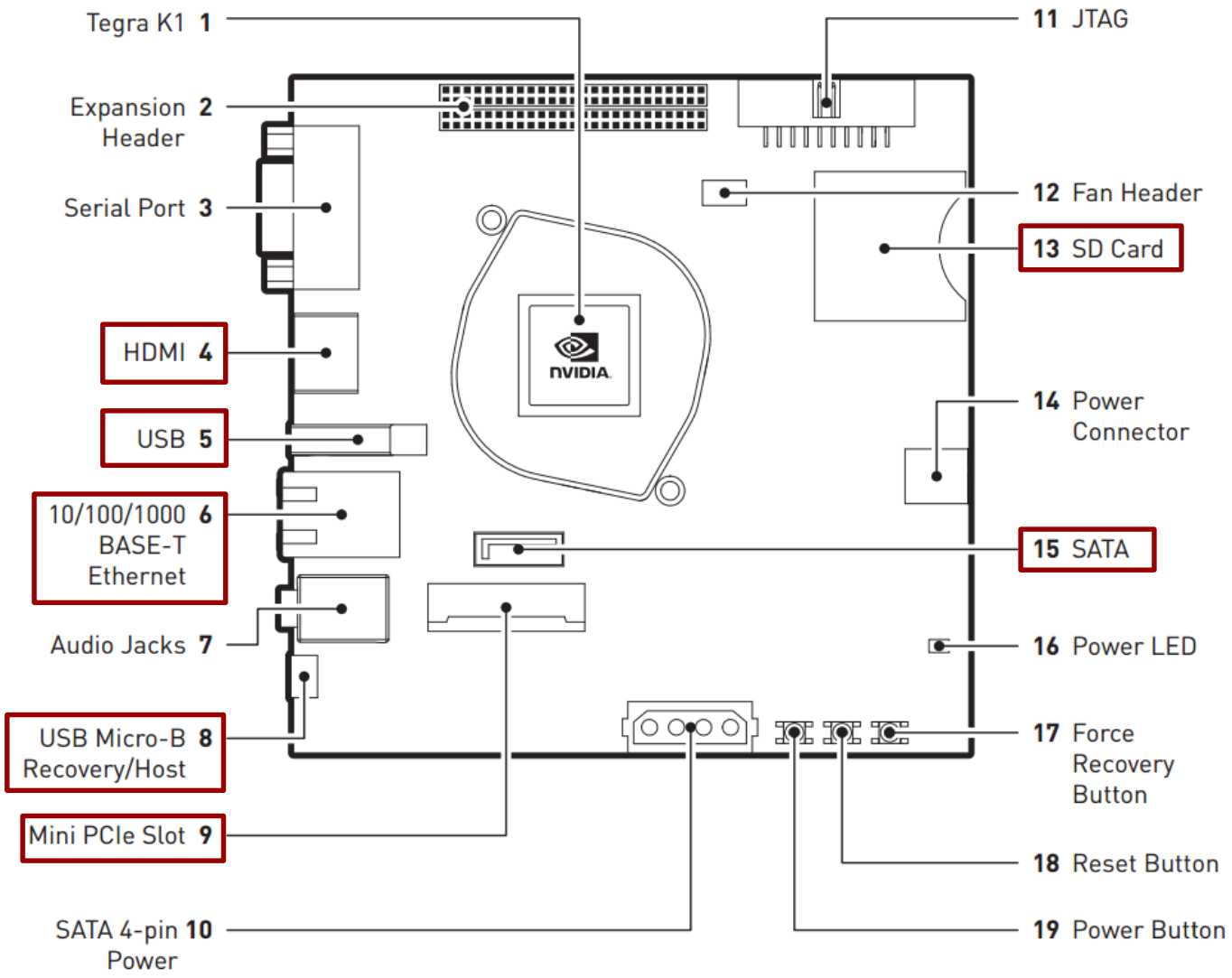
Tegra K1

- System on Chip (SOC)
- 4+1 cores ARM
- 192 cores Kepler
- CUDA
- OpenGL 4.4
- DirectX 11.1



src: http://on-demand.gputechconf.com/gtc/2014/presentations/S4412-tegra-k1-automotive-industry.pdf

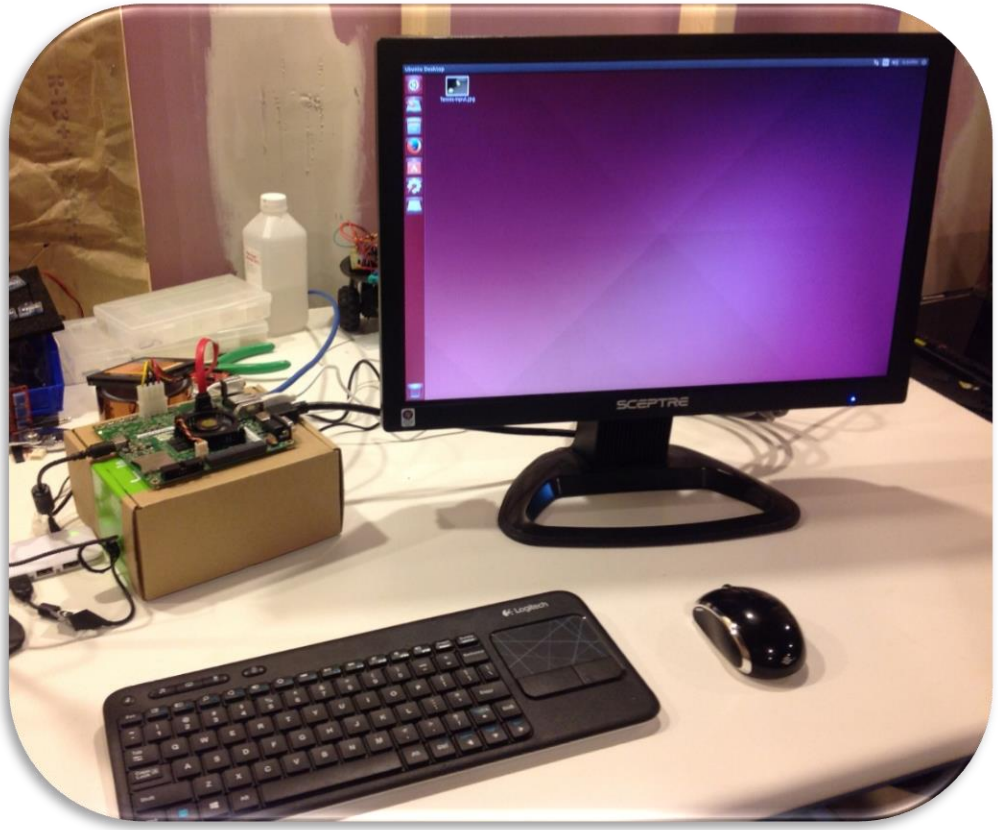
Jetson TK 1



src: http://developer.download.nvidia.com/embedded/jetson/TK1/docs/Jetson_TK1_QSG_134sq_Jun14_rev7.pdf

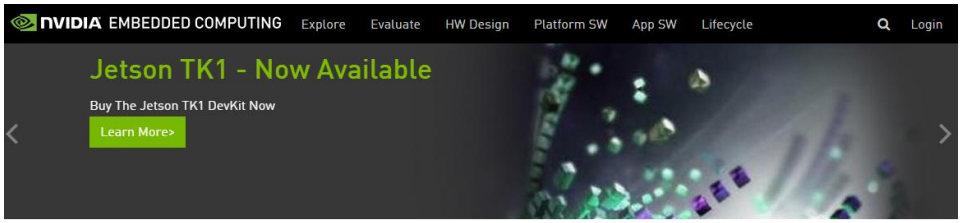
Jetson TK 1

- mini standalone computer
- Linux4Tegra (Ubuntu 14.04)
- CUDA Toolkit for L4T



src: http://secondrobotics.com/

Communities



Home > Embedded Computing

Meet the Jetson Embedded Platform
Learn more about our vision for embedded computing.

Getting Started with Jetson
Quick Start Guides, resources and the Jetson TK1 DevKit.

Hardware Design and Development
Design collateral, guidelines and tools to bring your project to life.

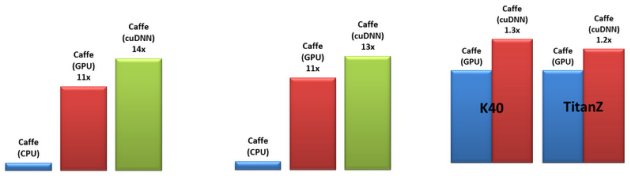
Platform Software Development
Board support packages (BSP), source code and documentation.

Application Software Development
Tools and resources to kickstart your application development.

Product Lifecycle Support
Production and support resources for your embedded project.

QUICKLINKS
- Jetson Wiki
- Embedded Forum
- Buy Jetson TK1 DevKit

cuDNN Performance Acceleration



Baseline Caffe compared to Caffe accelerated by cuDNN on K40

Baseline Caffe compared to Caffe accelerated by cuDNN on TitanZ

Baseline Caffe compared to Caffe accelerated by cuDNN

All comparisons are against a 12-core Intel E5-2679v2 CPU @ 2.4GHz running Caffe with Intel MKL 11.1.3.

Controlling GPU performance

To manually control the clock frequencies of the GPU, first determine the rates supported (listed by sysfs in kHz):

```
cat /sys/kernel/debug/clock/gbus/possible_rates
72000 108000 180000 252000 324000 396000 468000 540000 612000 648000 684000 708000 756000 804000 852000 (kHz)
```

Then set a rate (eg. the maximum of 852000kHz), specified in Hz:

```
echo 852000000 > /sys/kernel/debug/clock/override.gbus/rate
echo 1 > /sys/kernel/debug/clock/override.gbus/state
```

Finally verify the rate:

```
cat /sys/kernel/debug/clock/gbus/rate
852000
```

The gbus sysfs nodes control the GPU's core clock. To control the GPU's memory clock, substitute emc for gbus.

Main Page
Community portal
Current events
Recent changes
Help
Volunteering
Bug Tracker

Tools
What links here
Related changes
Special pages
Printable version
Permanent link
Page information

Page Discussion

Jetson TK1

About this site

This is the official Wiki for embedded Tegra & the Jetson TK1 board, maintained by both the community and NVIDIA.

The other embedded Tegra community sites with official NVIDIA support are:

- The forum for discussing embedded Tegra & Jetson TK1 issues with the community & NVIDIA.
- The blog to stay updated with the latest news & plans for embedded Tegra & Jetson TK1 from NVIDIA.

About Tegra K1

Tegra K1 is the world's first chip to have the same advanced features & architecture as a modern desktop GPU while still using the low power draw of a mobile chip! The Jetson TK1 board therefore allows embedded devices to use the exact same CUDA code that would also run on a desktop GPU (used by over 100,000 developers), with similar levels of GPU-accelerated performance as a desktop!

About Jetson TK1

Jetson TK1 is NVIDIA's embedded Linux development platform featuring a Tegra K1 SOC (CPU+GPU+ISP in a single chip), selling for $192 USD ($1 per CUDA core!). Jetson TK1 comes pre-installed with Linux4Tegra OS (basically Ubuntu 14.04 with pre-configured drivers). There is also some official support for running other distributions using the mainline kernel, discussed further in the Distributions and Mainline kernel sections below.

Besides the quad-core 2.3GHz ARM Cortex-A15 CPU and the revolutionary Tegra K1 GPU, the Jetson TK1 board includes similar features as a Raspberry Pi but also some PC-oriented features such as SATA, mini-PCIe and a fan to allow continuous operation under heavy workloads:

Hardware Features

- **Dimensions:** 5" x 5" (127mm x 127mm) board
- Tegra K1 SOC (CPU+GPU+ISP in a single chip, with typical power consumption between 1 to 5 Watts):
  - **GPU:** NVIDIA Kepler "GK20a" GPU with 192 SM3.2 CUDA cores (over 300 GFLOPS!)
  - **CPU:** NVIDIA "4-Plus-1" 2.32GHz ARM quad-core Cortex-A15 CPU with Cortex-A15 battery-saving shadow-core
- **DRAM:** 2GB DDR3L 933MHz EMC x16 using 64-bit data width

src: nvidia.com / elinux.org

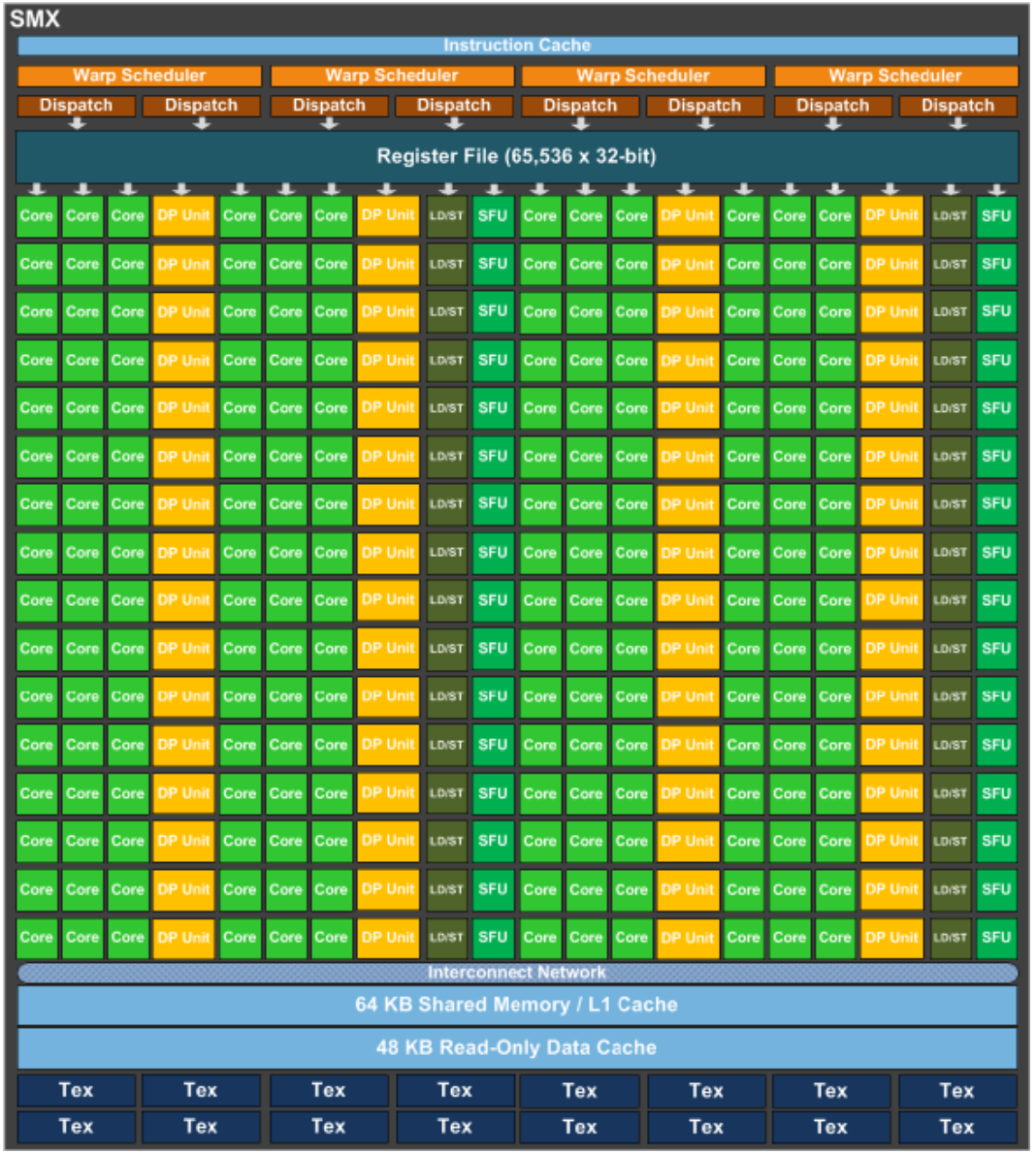
```
__global__ void adder(float *z, const float *x, const float *y) {
    int tid = blockIdx.x * blockDim.x + threadIdx.x;
    z[tid] = x[tid] + y[tid];
}

int main() {
    float *x = (float*) malloc(SIZE*sizeof(float));
    float *y = (float*) malloc(SIZE*sizeof(float));
    float *z = (float*) malloc(SIZE*sizeof(float));
    float *d_x, *d_y, *d_z;
    int i;

    for (i=0; i<SIZE; i++) {
        x[i] = i;
        y[i] = .1f * i;
    }
```

GPU Basics

src: www.nvidia.com

GPU Architecture

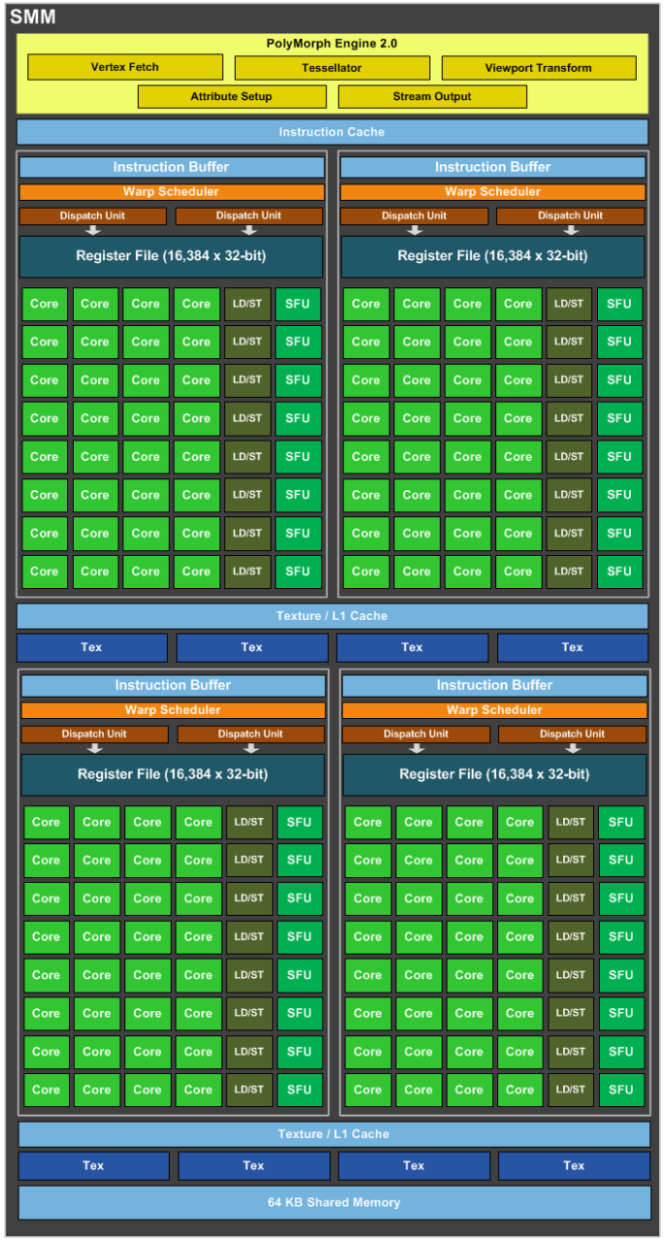
- Kepler SMX
  192 Cores
- Four Schedulers
- 64 KB Shared Memory



src: http://www.nvidia.com/content/PDF/kepler/NVIDIA-kepler-GK110-Architecture-Whitepaper.pdf

GPU Architecture

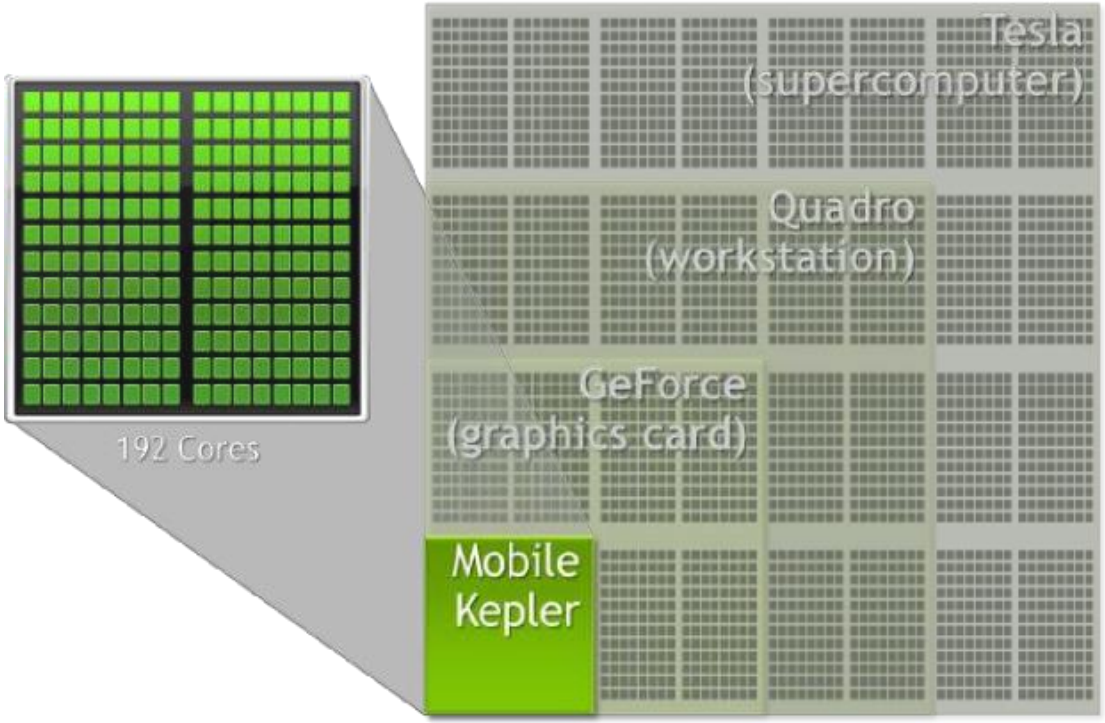
- Maxwell SMM – 128 Cores
- Four Schedulers
- 64 KB Shared Memory



src: anandtech.com

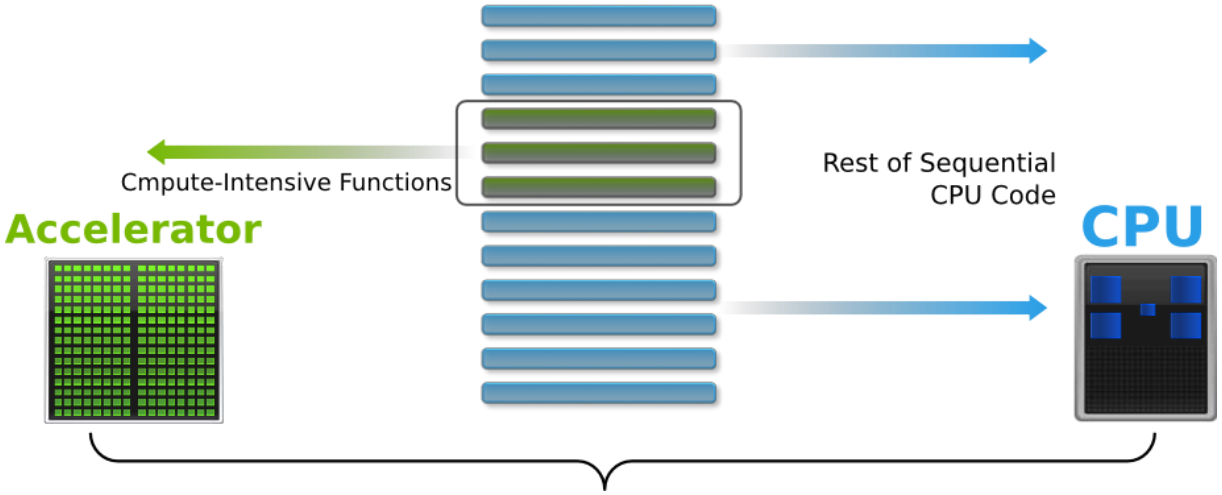
Tegra, GeForce, Quadro and Tesla

- Tegra K1
  - 192 CUDA cores
- GeForce GT740 (GK107)
  - 384 CUDA cores
- Quadro K4200 (GK104)
  - 1344 CUDA cores
- Tesla K20m
  - 2496 CUDA cores



src: http://on-demand.gputechconf.com/gtc/2014/presentations/S4906-mobile-compute-tegra-K1.pdf

GPU Basics

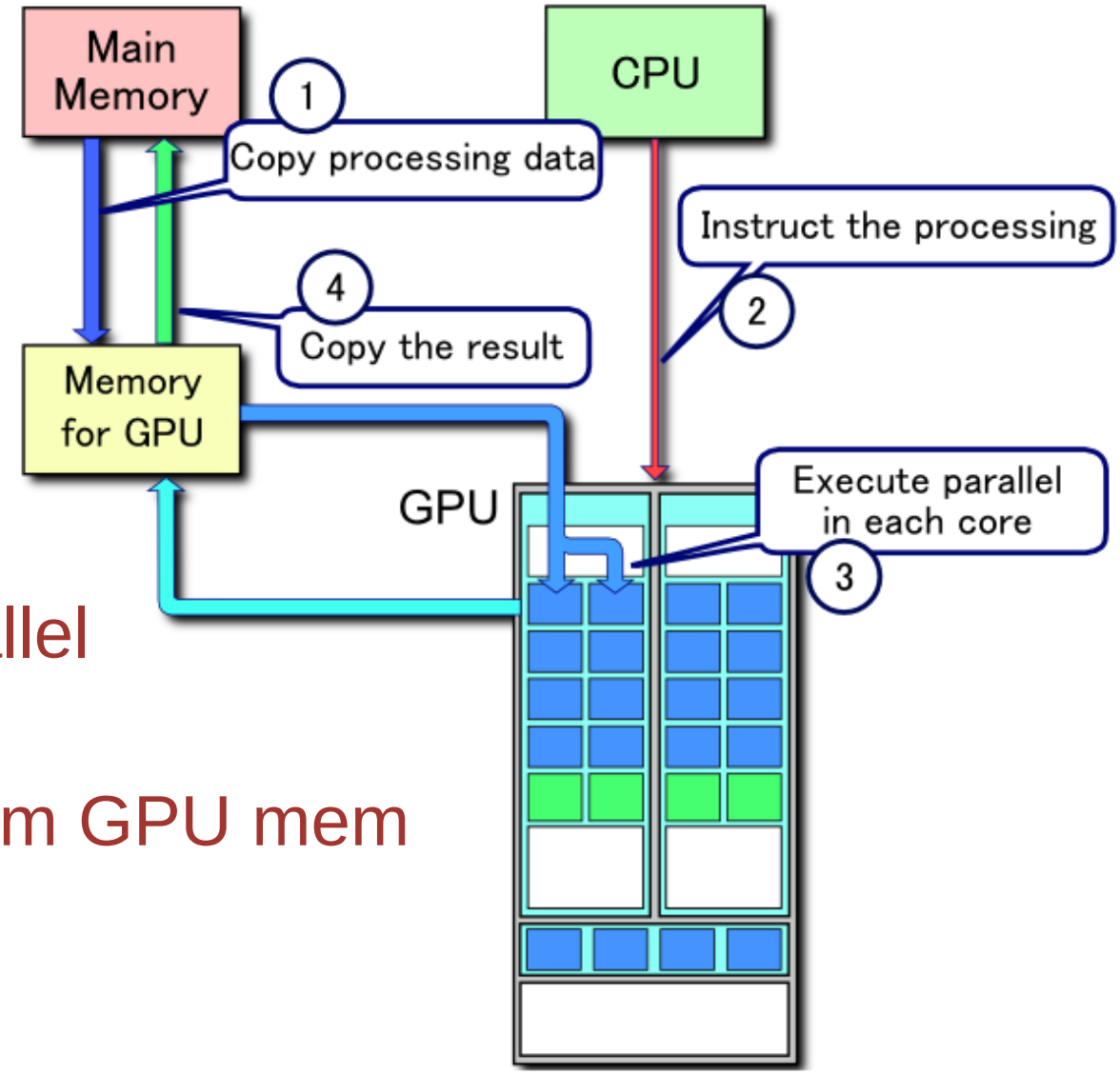


- Why GPUs?
  - High throughput of parallel workloads
- Workload has to be divided in serial and parallel Sections

src: http://on-demand.gputechconf.com/gtc/2014/presentations/S4906-mobile-compute-tegra-K1.pdf

Processing flow for GPU transfers

- Copy data from main mem to GPU mem
- CPU instructs the process to GPU
- GPU execute parallel in each core
- Copy the result from GPU mem to main mem



src: http://upload.wikimedia.org/wikipedia/commons/5/59/CUDA_processing_flow_%28En%29.PNG

SAXPY serial and SAXPY parallel

```
void saxpy_serial(int n, float a, float *x, float *y)
{
   for (int i = 0; i < n; ++i)
      y[i] = a*x[i] + y[i];
}
```

```
__global__ void saxpy_parallel(int n, float a, float *x, float *y)
{
   int i = blockIdx.x*blockDim.x + threadIdx.x;
   if (i < n)  y[i] = a*x[i] + y[i];
}
```

- For-Loop now in parallel
- BlockID and ThreadID identify the threads

src: http://gpulab.compute.dtu.dk/PhDschool/slides/CUDA%20Tutorial.pdf

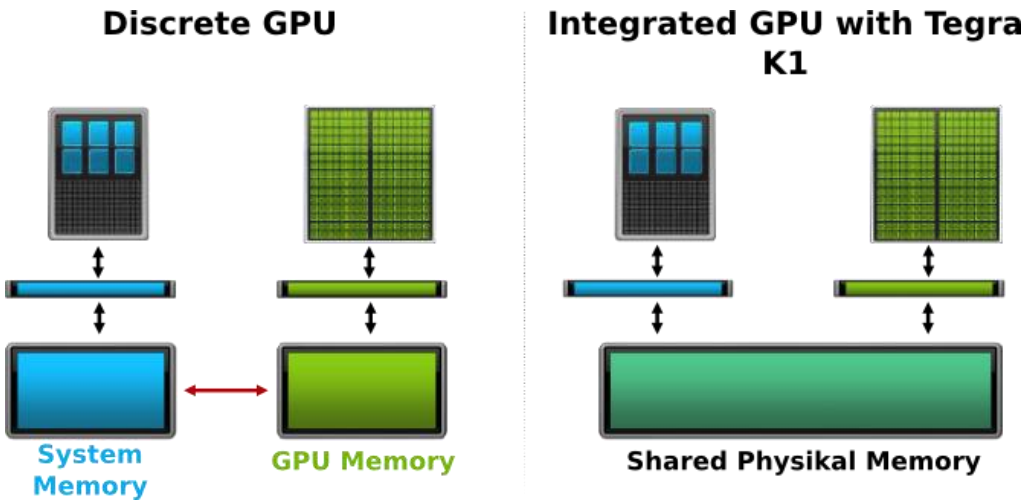
SAXPY: Host Code

- cudaMalloc – allocate memory on the device
- cudaMemcpy – copy data between host and device
  - HostToDevice
  - DeviceToHost
- <<< … >>> - # of blocks and threads per block

```
// Allocate two N-vectors h_x and h_y
int size = N * sizeof(float);
float* h_x = (float*)malloc(size);
float* h_y = (float*)malloc(size);

// Initialize them...

// Allocate device memory
float* d_x; float* d_y;
cudaMalloc((void**)&d_x, size));
cudaMalloc((void**)&d_y, size));

// Copy host memory to device memory
cudaMemcpy(d_x, h_x, size, cudaMemcpyHostToDevice);
cudaMemcpy(d_y, h_y, size, cudaMemcpyHostToDevice);

// Invoke parallel SAXPY kernel with 256 threads/block
int nblocks = (N + 255) / 256;
saxpy_parallel<<<nblocks, 256>>>(N, 2.0, d_x, d_y);

// Copy result back from device memory to host memory
cudaMemcpy(h_y, d_y, size, cudaMemcpyDeviceToHost);
```

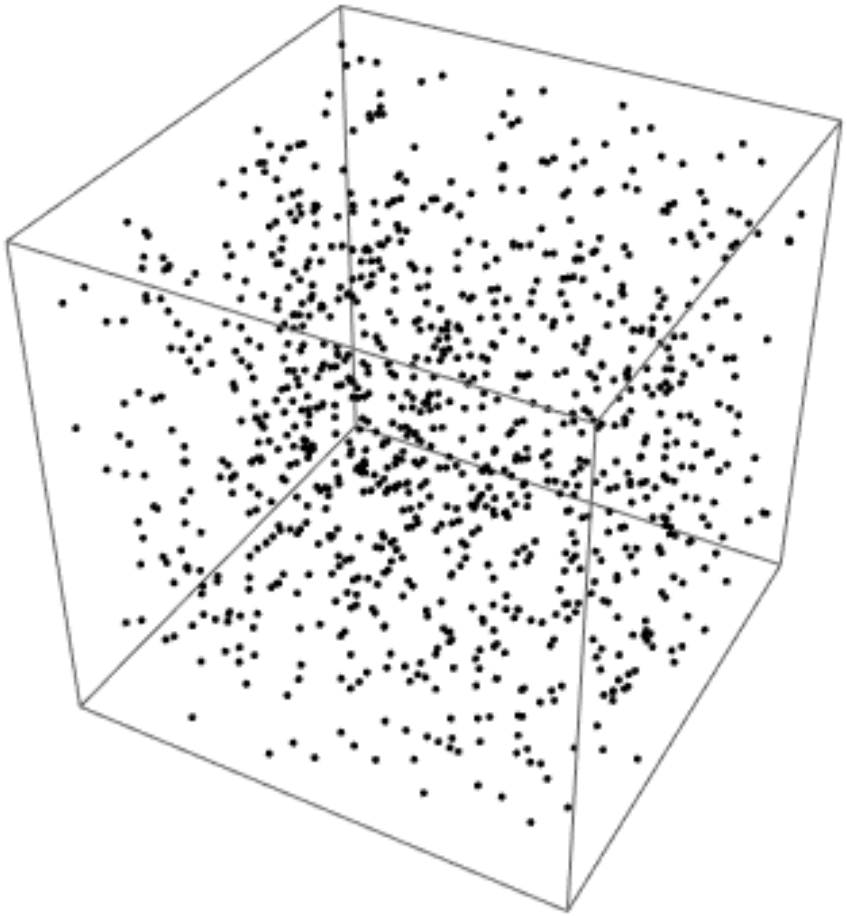
src: http://gpulab.compute.dtu.dk/PhDschool/slides/CUDA%20Tutorial.pdf

Shared Physikal Memory



- No communication overheads
- No cudaMemcpy
- caching benefits

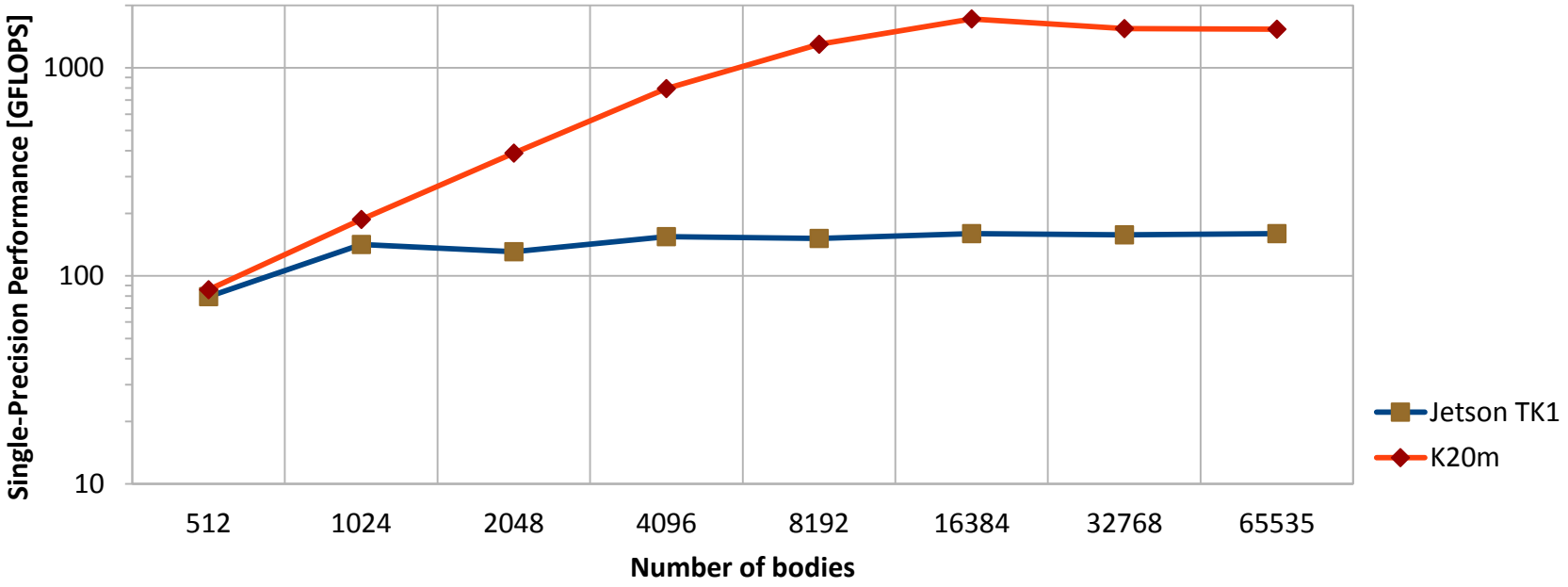
src: http://on-demand.gputechconf.com/gtc/2014/presentations/S4906-mobile-compute-tegra-K1.pdf



Benchmark

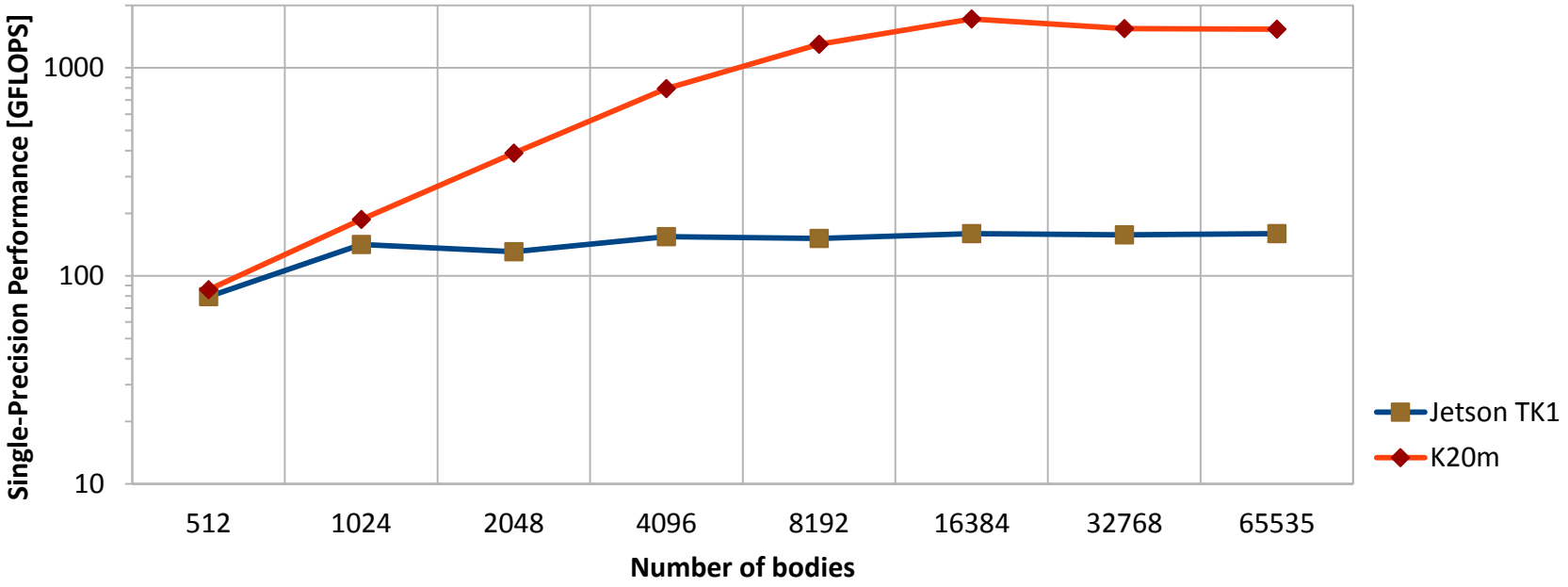
src: http://community.wolfram.com/groups/-/m/t/173763

nBody Benchmark



- K20m:
  - 2x Intel Ivy Bridge E5-2630 – 2.6 GHz
  - 64 GB RAM
  - Tesla K20m – 2496 CUDA cores
- Jetson TK1:
  - 4x ARM Cortex A15
  - 2 GB RAM
  - GK20a – 192 CUDA cores

nBody Benchmark



Number of Bodies	Jetson TK1 [GFLOPS]	K20m [GFLOPS]
512	79,478	85,902
1024	141,859	186,691
2048	130,971	389,788
4096	154,432	794,556
8192	151,609	1300,601
16384	159,609	1721,291
32768	157,642	1547,459
65535	159,852	1535,320

Power Efficiency

Power Efficiency

System Status	Power [W] SP	GFlops SP	GFlops/W SP	Power [W] DP	GFlops DP	GFlops/W DP
boot	up to 6.5	-	-	-	-	-
idle	3.2	-	-	-	-	-
nBody (energy saving)	4.2	13.4	3.2	3.8	0.9	0.23
nBody (GPU max clock rate)	14.2	159.9	11.3	7.4	10.9	1.5
nBody on K20m (only GPU)	162	1753	10.8	153	596.4	3.9

Green500 Rank	Mflops/Watt	Name	Computer
1	5271,8142	L-CSC	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150
2	4945,625592	Suiren	ExaScaler 32U256SC Cluster, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, PEZY-SC
3	4447,584063	TSUBAME-KFC	LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x
4	3962,73013	Storm1	Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40m
5	3631,864623	Wilkes	Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20
6	3543,315018		iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x
7	3517,83674	HA-PACS TCA	Cray CS300 Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x
8	3459,459459	Cartesius Accelerator Island	Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4× FDR, Nvidia K40m
9	3185,908329	Piz Daint	Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x
10	3131,06498	romeo	Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x

src: http://www.green500.org/

Power Efficiency

System Status	Power [W] SP	GFlops SP	GFlops/W SP	Power [W] DP	GFlops DP	GFlops/W DP
boot	up to 6.5	-	-	-	-	-
idle	3.2	-	-	-	-	-
nBody (energy saving)	4.2	13.4	3.2	3.8	0.9	0.23
nBody (GPU max clock rate)	14.2	159.9	11.3	7.4	10.9	1.5
nBody on K20m (only GPU)	162	1753	10.8	153	596.4	3.9

Green500 Rank	Mflops/Watt	Name	Computer
1	5271,8142	L-CSC	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150
2	4945,625592	Suiren	ExaScaler 32U256SC Cluster, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, PEZY-SC
3	4447,584063	TSUBAME-KFC	LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x
4	3962,73013	Storm1	Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40m
5	3631,864623	Wil	Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20
6	3543,315018		iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x
7	3517,83674	HA-PACS TCA	Cray CS300 Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x
8	3459,459459	Cartesius Accelerator Island	Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4× FDR, Nvidia K40m
9	3185,908329	Piz Daint	Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x
10	3131,06498	romeo	Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x

src: http://www.green500.org/

Power Efficiency

System Status	Power [W] SP	GFlops SP	GFlops/W SP	Power [W] DP	GFlops DP	GFlops/W DP
boot	up to 6					
idle	3.2					
nBody (energy saving)	4.2	13.4	3.2	3.8	0.9	0.23
nBody (GPU max clock rate)	14.2	159.9	11.3	7.4	10.9	1.5
nBody on K20m (only GPU)	162	1753	10.8	153	596.4	3.9

single precision

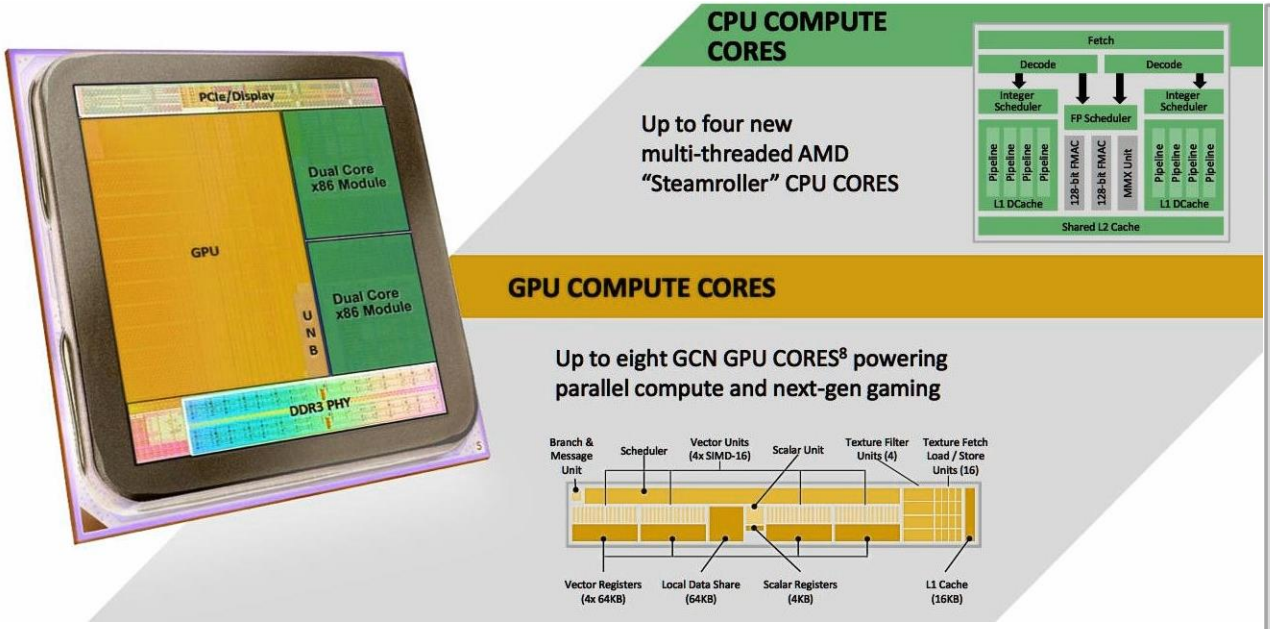
double precision

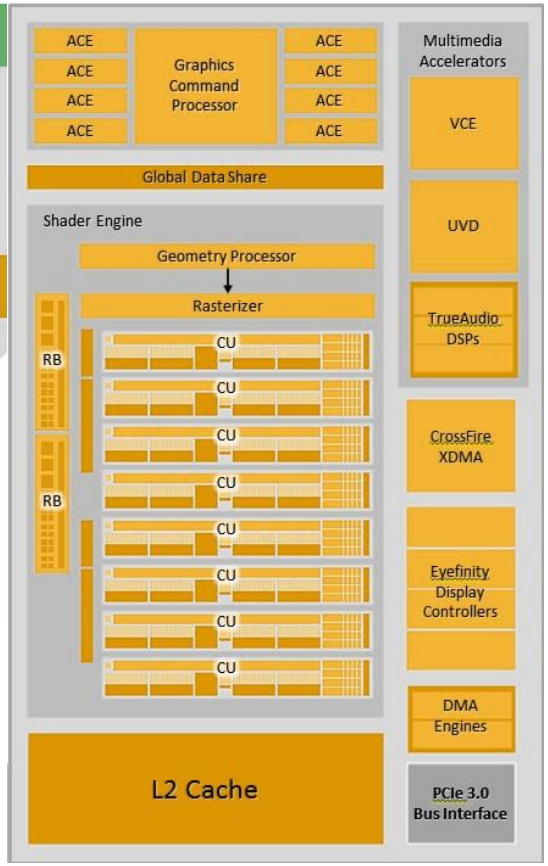
Green500 Rank	Mflops/Watt	Name	Computer
1	5271,8142	L-CSC	SC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150
2	4945,625592	Suiren	Exa er 32U256SC Cluster, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, PEZY-SC
3	4447,584063	TSUBAME-KFC	LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x
4	3962,73013	Storm1	Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40m
5	3631,864623	Wil	Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20
6	3543,315018		iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x
7	3517,83674	HA-PACS TCA	Cray CS300 Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x
8	3459,459459	Cartesius Accelerator Island	Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4× FDR, Nvidia K40m
9	3185,908329	Piz Daint	Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x
10	3131,06498	romeo	Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x

double precision

src: http://www.green500.org/

Related Work





- AMD APU (Kaveri A10-7800):
  - 12 Compute Cores (4 CPU + 8 GPU)
  - 512 Shader Arithmetic Units (8 x 64)
- AMD APU (Temash A6-1450):
  - 6 Compute Cores (4 CPU + 2 GPU)
  - 128 Shader Arithmetic Units (2 x 64)

src: hksilicon.com

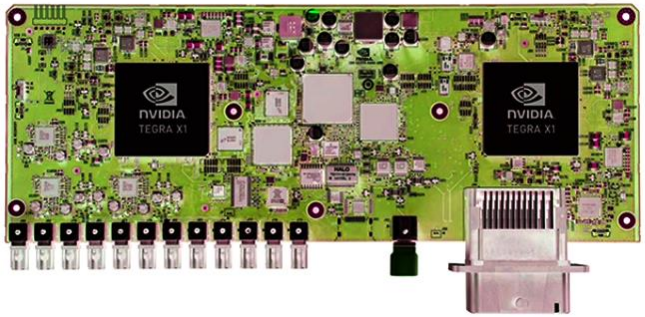
Future – Tegra X1

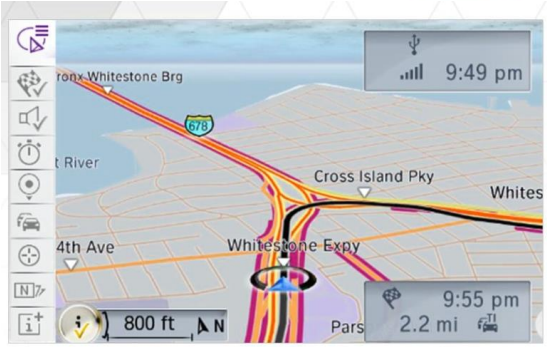
INTRODUCING NVIDIA DRIVE™ PX

AUTO-PILOT CAR COMPUTER

Dual Tegra X1 • 12 camera inputs • 1.3 GPix/sec

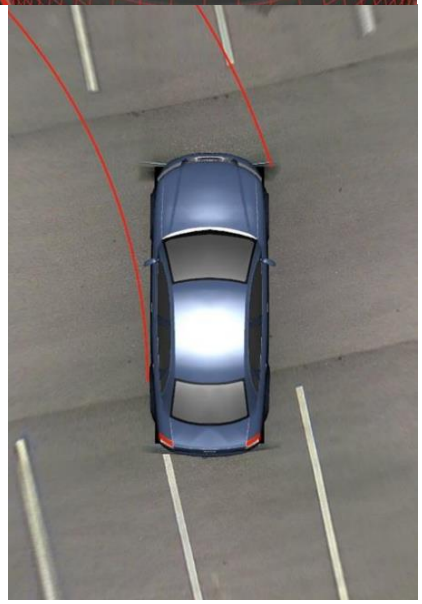
- 2.3 Teraflops mobile supercomputer
- CUDA programmability
- Deep Neural Network Computer Vision
- Surround Vision





TODAY

DRIVE CX



src: http://international.download.nvidia.com/pdf/tegra/Tegra-X1-whitepaper-v1.0.pdf

Future – Mont-Blanc

- setting future global HPC standards
- solutions used in embedded and mobile devices
- support for ARMv8 64-bit processors

MONT-BLANC
EUROPEAN APPROACH TOWARDS ENERGY EFFICIENT HIGH PERFORMANCE

ODROID-XU

24 nodes, each node is equiped with:

- SoC Samsung Exynos 5 Octa 5410
- CPU Cortex-A15@1.6GHz quad core and Cortex-A7@1.2GHz quad core
- GPU PowerVR SGX544MP3 - No OpenCL support available
- 2Gbyte LPDDR3 RAM PoP
- 1 Gb Ethernet interconnection

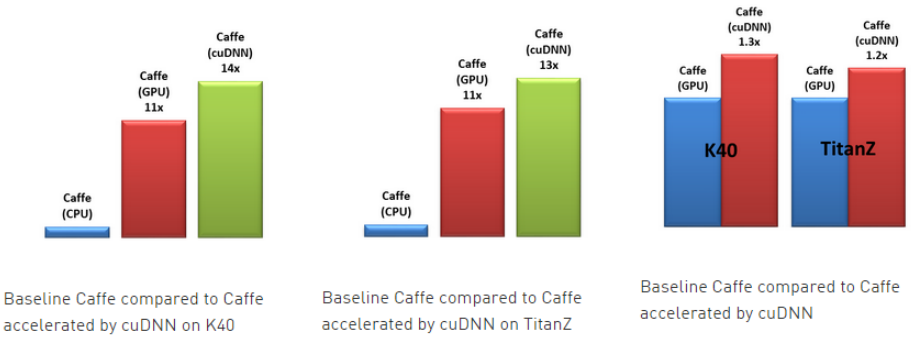
JETSON TK1

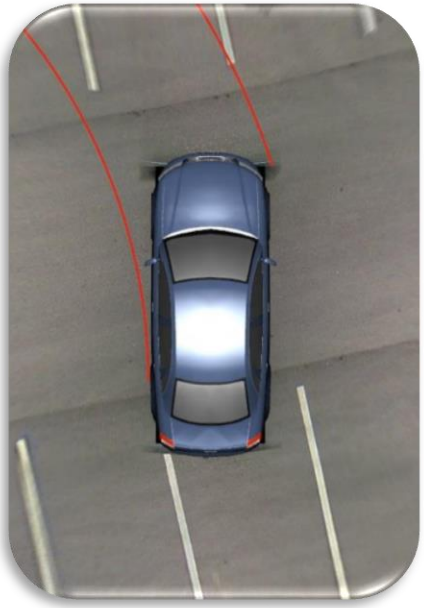
8 nodes, each equipped with:

- Nvidia Tegra K1 SoC
- 4-Plus-1 quad-core ARM Cortex A-15 CPU (4x Cortex-A15 + 1x Cortex-A7)
- Kepler GPU with 192 cores
- 2 Gbyte memory with 64 bit width

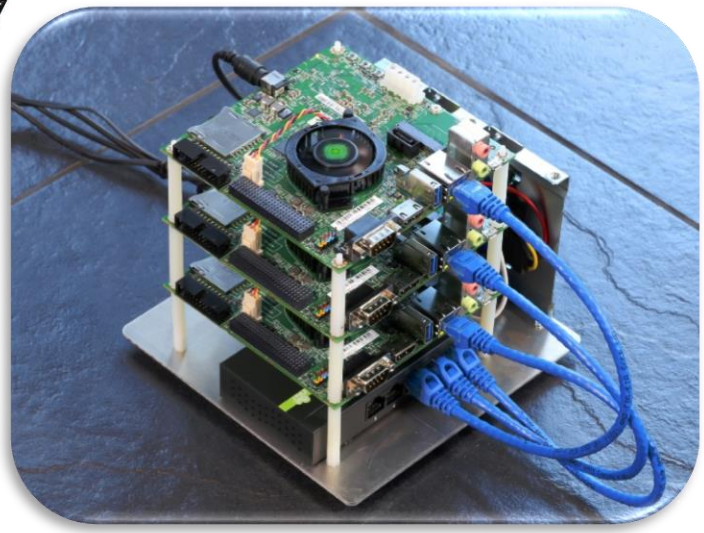
src: http://montblanc-project.eu/

Conclusion





- Robots with deep neuronal networks
- Energy efficient Supercomputer
- Saver and more comfortable Vehicles



src: elinux.org / nvidia.com