電気通信大学

UEC TOKYO

The University of Electro-communications, Tokyo

“High Performance Computing on Mobile Devices through Distributed Shared CUDA”

By

Martinez Noriega Edgar Josafat.
Dr. Narumi Tetsu.

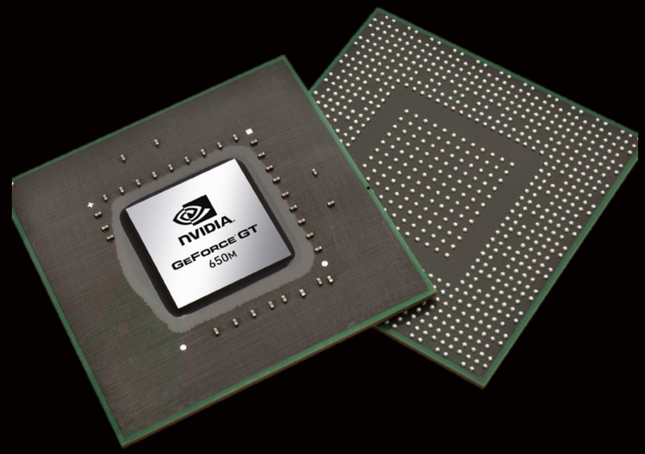
Introduction

GPUs are everywhere!

GPU - Graphics Processor Unit

GPU characteristics:

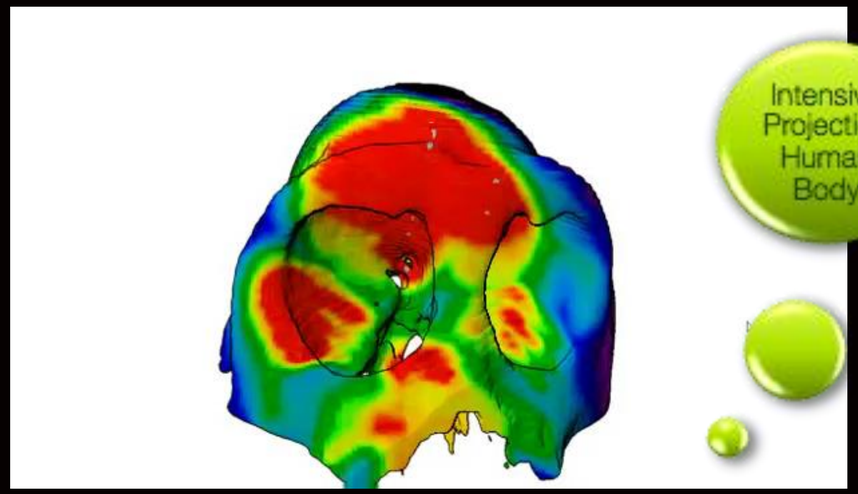
- ➡Massively programable parallel processors.
- ➡Different memory hierarchy.
- ➡Multithreads many core chips.



Advantages:

- ➡Very attractive performance/cost benefit.
- ➡Multipurpose e.g. Gaming, GPGPU, Rendering

HPC - Applications



CUDA Applications

Intensive Projection Human Body

Galaxy Collision

Milky Way - Andromeda Galaxy Collision

Molecular Dynamics

Atacking H1N1 Virus

Mobile Devices

- Mobility.
- Portability.
- Connectivity.
- Huge ecosystem.
- Limited memory.
- Low power consumption.
- Low processing (ARM processors)
- Touch screen capabilities

Merging Mobile Devices and HPC apps

Where to get such acceleration ?

How to get such acceleration ?

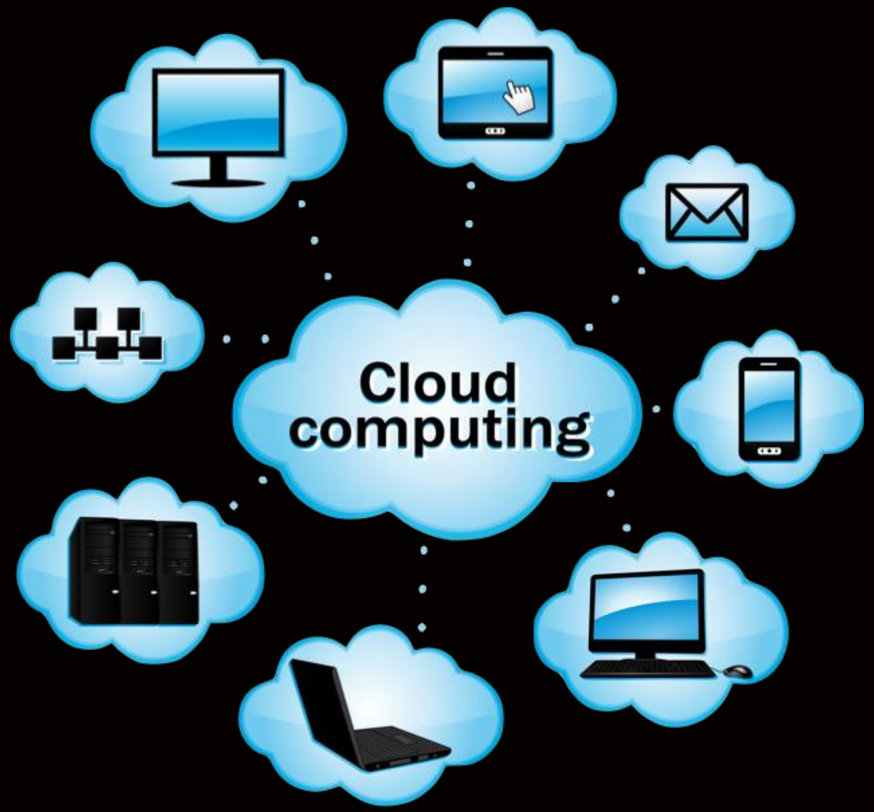
When to get such acceleration ?

Cloud Computing

- Cloud computing is promising since the user can use arbitrary computing power on demand from anywhere.

Examples:

- Amazon EC2 (Elastic Compute Cloud)
- IBM Computing on Demand
- NVIDIA VGX
- NVIDIA GeForceGRID

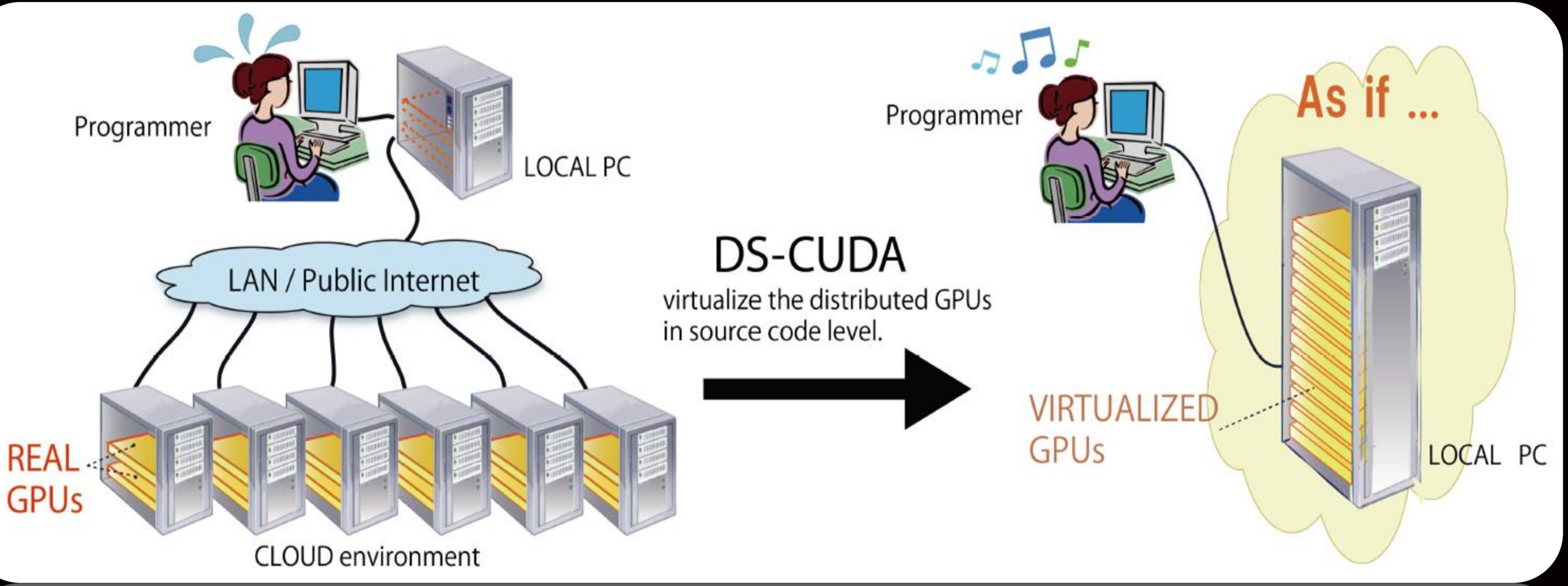


GPU virtualization software

- DS-CUDA = Distributed Shared Compute Unified Device Architecture
- DS-CUDA is open source. http://narumi.cs.uec.ac.jp/dscuda/
- Middleware to simplify the development of code that uses multiple GPUs.
- It virtualizes a cluster of GPUs equipped PCs to seem like a single PC with many GPUs.
- The perfomance of Many Body simulation has been tested on 22-node (64-GPU) TSUBAME 2.0 supercomputer.

*Atsushi Kawai, Kenji Yasuoka, Kazuyuki Yoshikawa and Narumi Tetsu “Distributed Sahred CUDA:Virtualization of Large-Scale GPU systems for Pragammability and Reliability”The Fourth International Conference on Future Computational Technologies and Applications, France 2012)

DS-CUDA system overview.



DS-CUDA Package contents

Server:

- Server daemon
- ./dscudaserver
- Configurable by Env. Variables: export DSCUDA_WARNLEVEL=5
- Source code

Client:

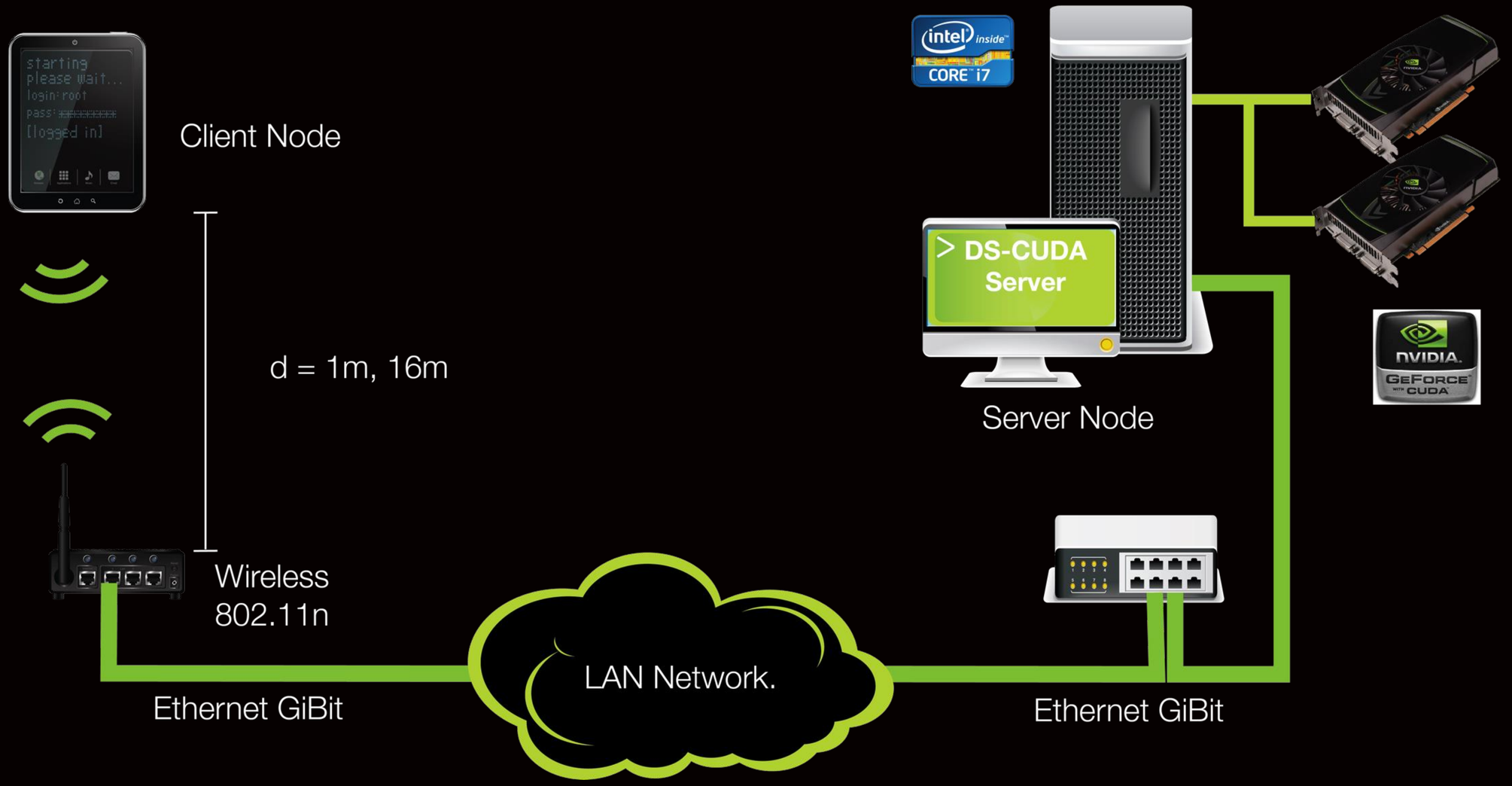
- Compiler
- SDK (Matrixmul, Vecadd, Claret, Bandwidth_test, MultiGPU,etc
- Configurable by Env. Variables: export DSCUDA_SERVER= 192.168.0.110
- Source code

GPU virtualization software

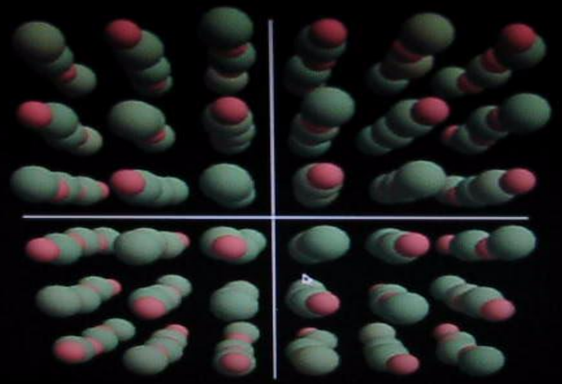
DS-CUDA main specifications.

Spec	Client	Server
Network	RPC(Socket) InfiniBand (Verb)	RCP (Socket) InfiniBand (Verb)
Architecture OS	64 bit	64 bit
Host OS	Linux	Linux
CUDA		4.2

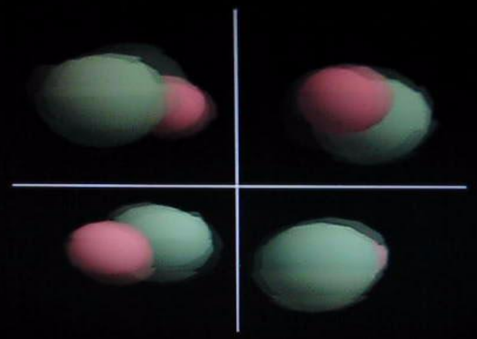
System Architecture: DS-CUDA-Tablet



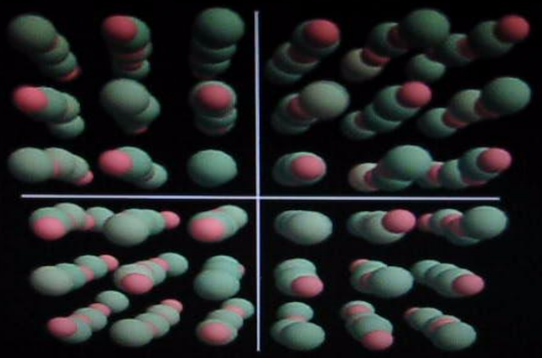
Molecular Dynamics Simulation - Claret



Shot 27 new ions



Number of Particles:
{8, 64, 216, 512, 1000, 1728, 2744, 4096, 5832}

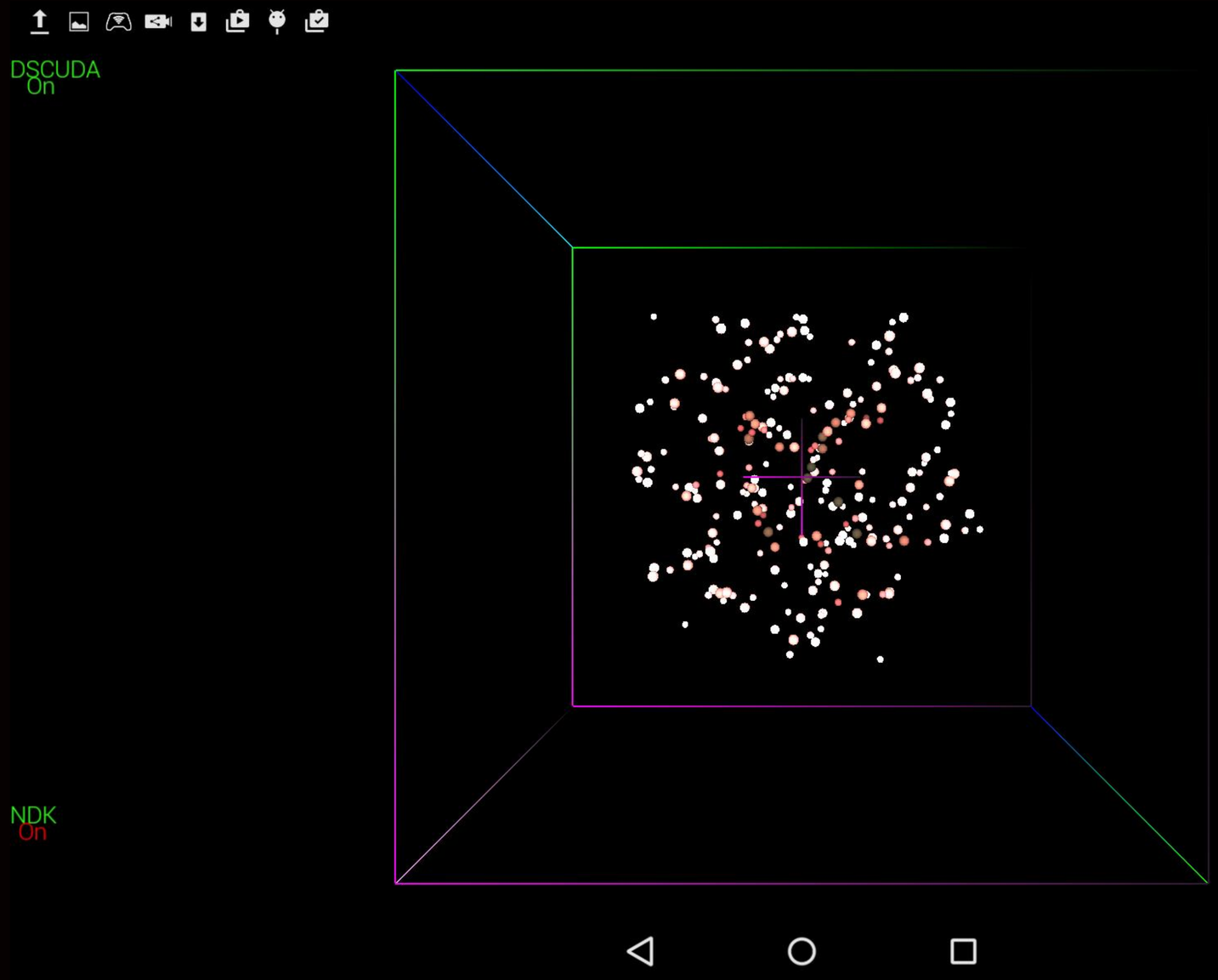


Graphical Detail

Characteristics of CS:

- CS is a scientific data visualizations tool created by Dr. Takahiro Koishi on 2001
- Emulates and presents (through graphics) the behavior between NaCl particles at vacuum level.
- Computes the Force between NaCl particles (Tosi-Fumi method)
- Positions and velocities of atoms are updated by Newton's equation of motion (Time integration).
- Source code in C language and open graphics library (OpenGL) for visualization part.

Molecular Dynamics on Tablet



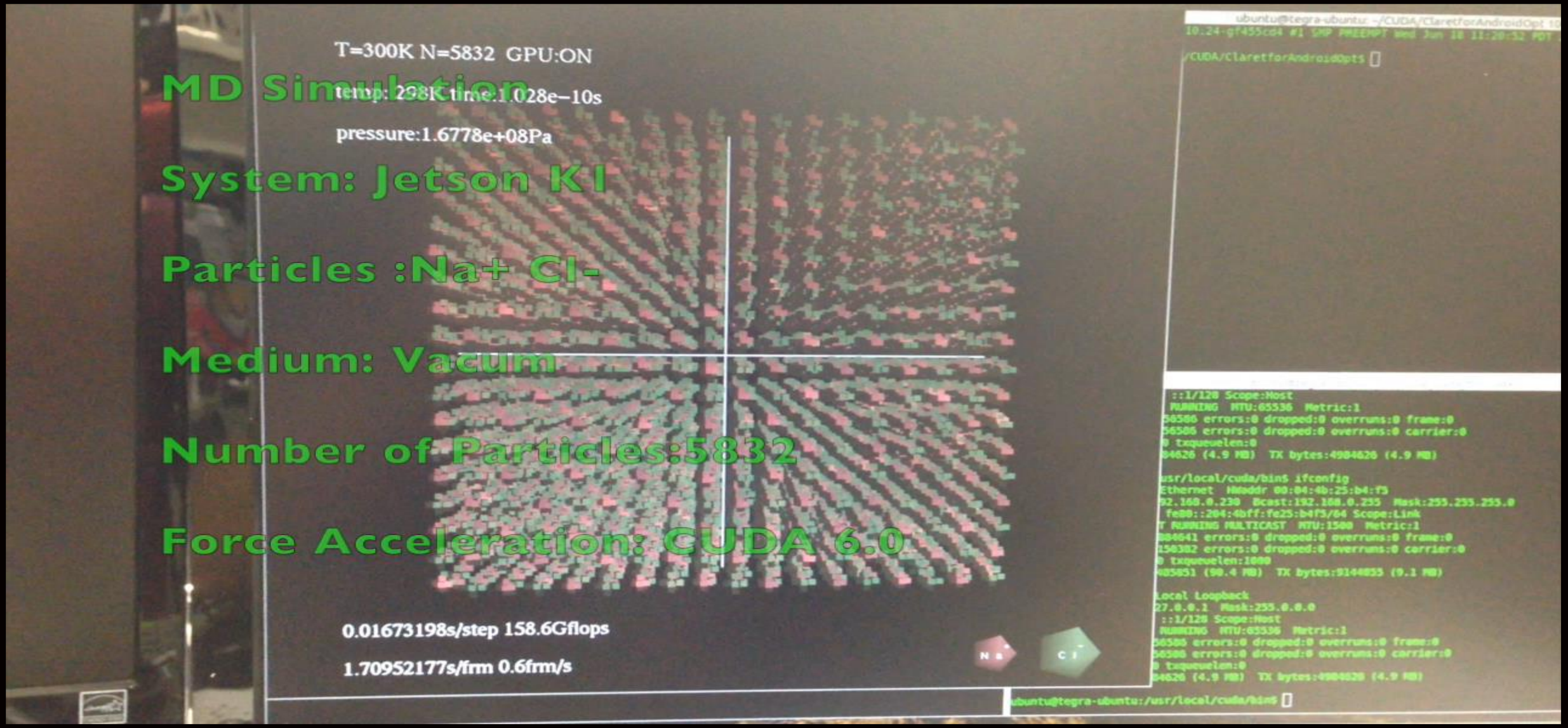
- Multi Gestures Enable
  - 1 Finger - Rotate
  - 2 Fingers - Zoom
  - 3 Fingers - Perspective
- Switching Force Calculation medium Enable
  - DS-CUDA - Remote GPU
  - ARM - CPU
- Flops Performance information Enable
- Shoot New 27 Ions

System test: Characteristics

Machines Test Specifications

Device	CPU	GPU	Memory	OS	CUDA
Alienware Knoppix 7.02 32	Intel Core i7, 2.30 GHz, 8 Cores	GeForce GT 680M, 7 MultiProcessors, 1344 CUDA Cores, Global Memory 2047Mbytes.	16 Gbytes, DDR3, 1600 MHz	Knoppix7.0.2 x86 Linux	Driver 331.62, Toolkit 6.0, SDK 6.0
NVIDIA “SHIELD”	NVIDIA Tegra 4, ARMv7, 1.912 GHz, 4 Cores	NVIDIA AP, 72 Custom Cores,	2 Gbytes, DDR3L & LPDDR3	Android 4.4.2	—
Tegra K1	Intel Core i7, 2.40 GHz, 8 Cores	Tegra K1 (GK20A), 1 MultiProcessors, 192 CUDA Cores, Global Memory 1746 Mbytes.	2 Gbytes, DDR3L, 933 MHz	Linux for Tegra (Ubuntu 14.04 for ARM)	Driver “Custom for Jetson K1”, Toolkit 6.0, SDK 6.0

Demo


T=300K N=5832 GPU:ON
MD Simulation
temp: 298K time:1.028e-10s
pressure:1.6778e+08Pa
System: Jetson KI
Particles :Na+ Cl-
Medium: Vacuum
Number of Particles:5832
Force Acceleration: CUDA 6.0
0.01673198s/step 158.6Gflops
1.70952177s/frm 0.6frm/s
Na+
Cl-

DS-CUDA on Android

Porting DS-CUDA (client) to Android - Challenges:

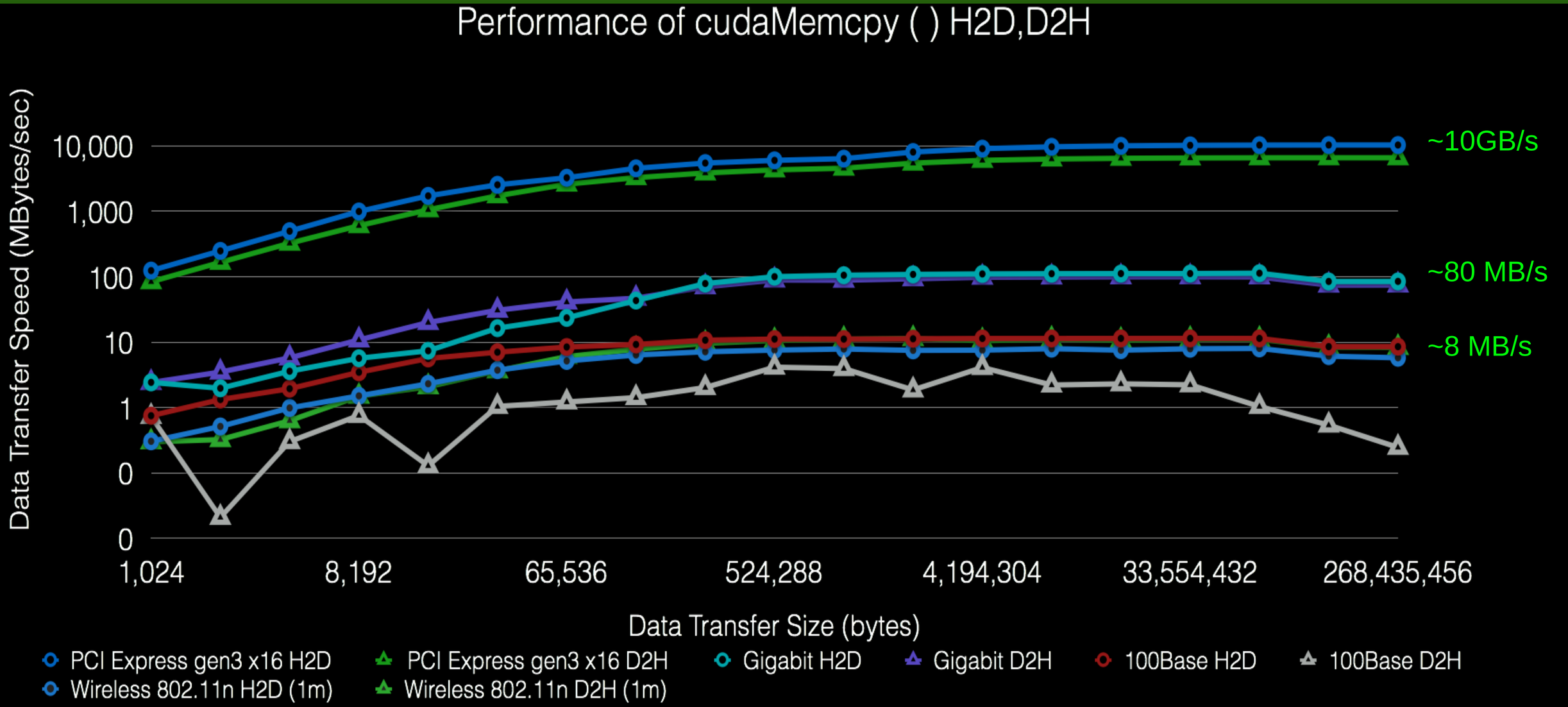
➡RPC (Remote Procedure Call) is not supported on Android
  ➡Used only TCP socket

➡ C/C++ code loading inside of Java code
  ➡Use NDK (Native Development Kit) to generate DS-CUDA code inside of static library.

➡ 64-bit DS-CUDA server cannot be used
  ➡Modified the server to work in 32-bit (Linux/Knoppix).

➡Differences in searching host name in socket API
  ➡Change the hand shaking and retrieval information. Before was RPC.

Bandwidth between different mediums.



"Bandwidth Test" sample from CUDA SDK is used.

Model of MD simulator for Analysis.

$$T = T\_GPU + T\_CPU + T\_COMM + T\_DISP$$

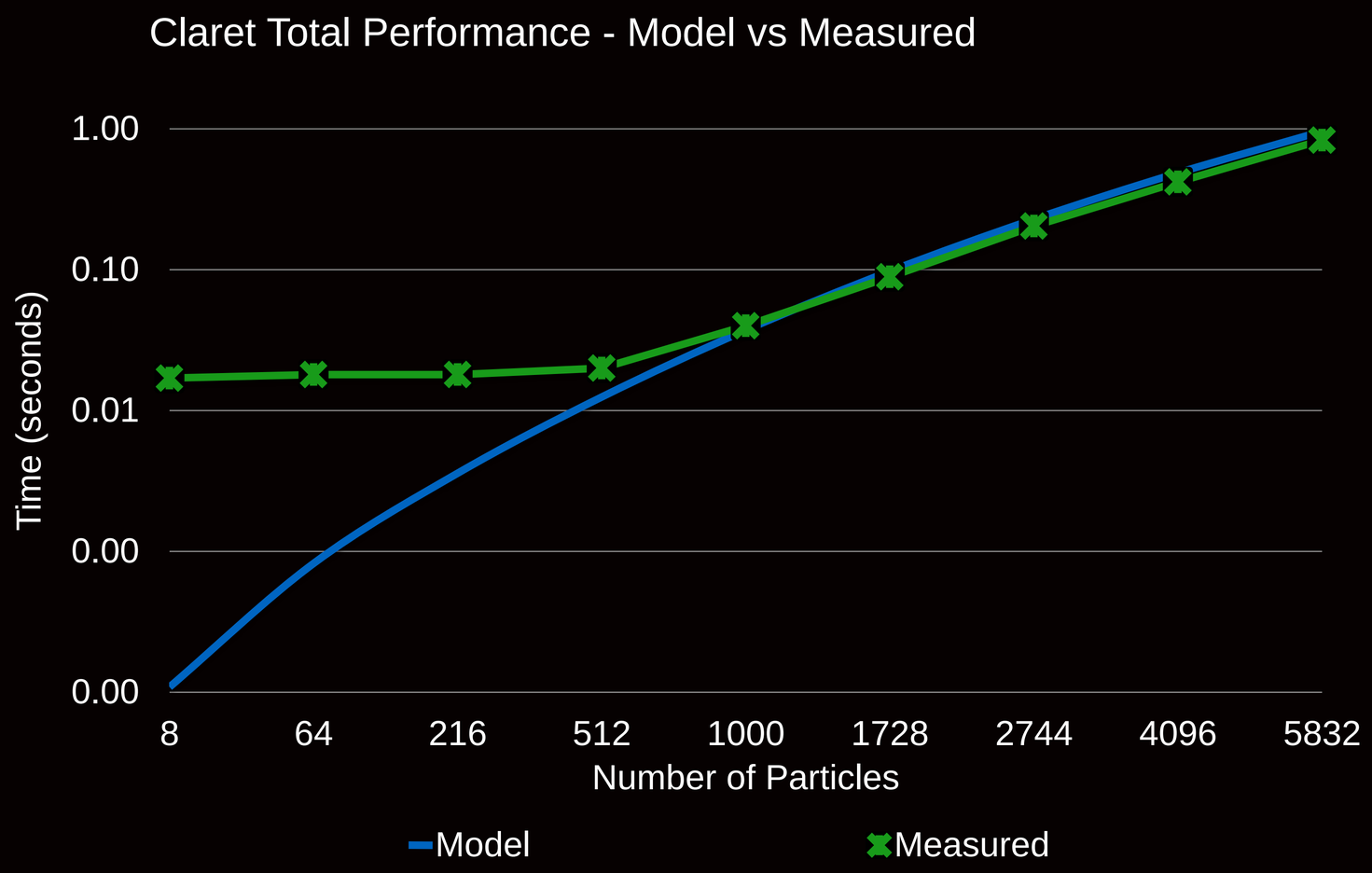
T :Time per Frame on Claret Demo

T_GPU: Time on GPU

T_CPU: Time onCPU

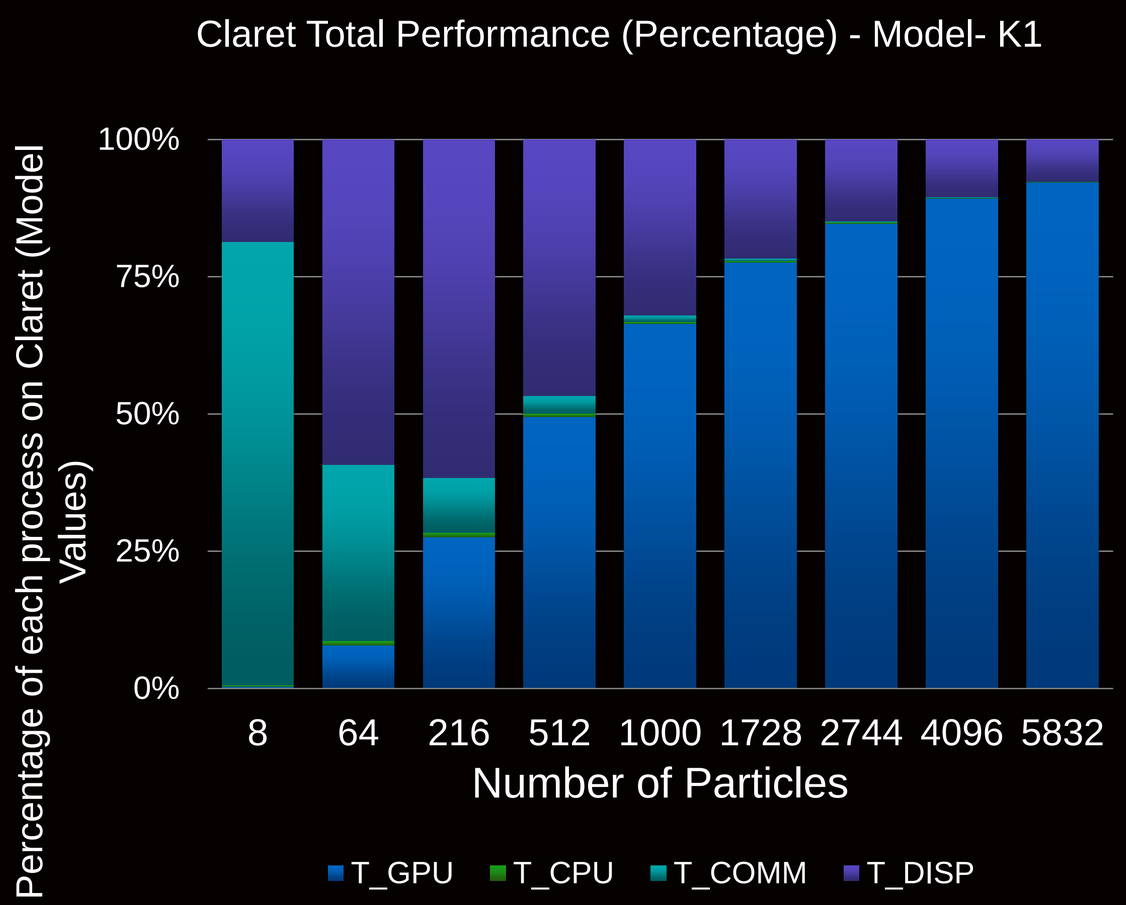
T_COM: Time for communication between CPU and GPU

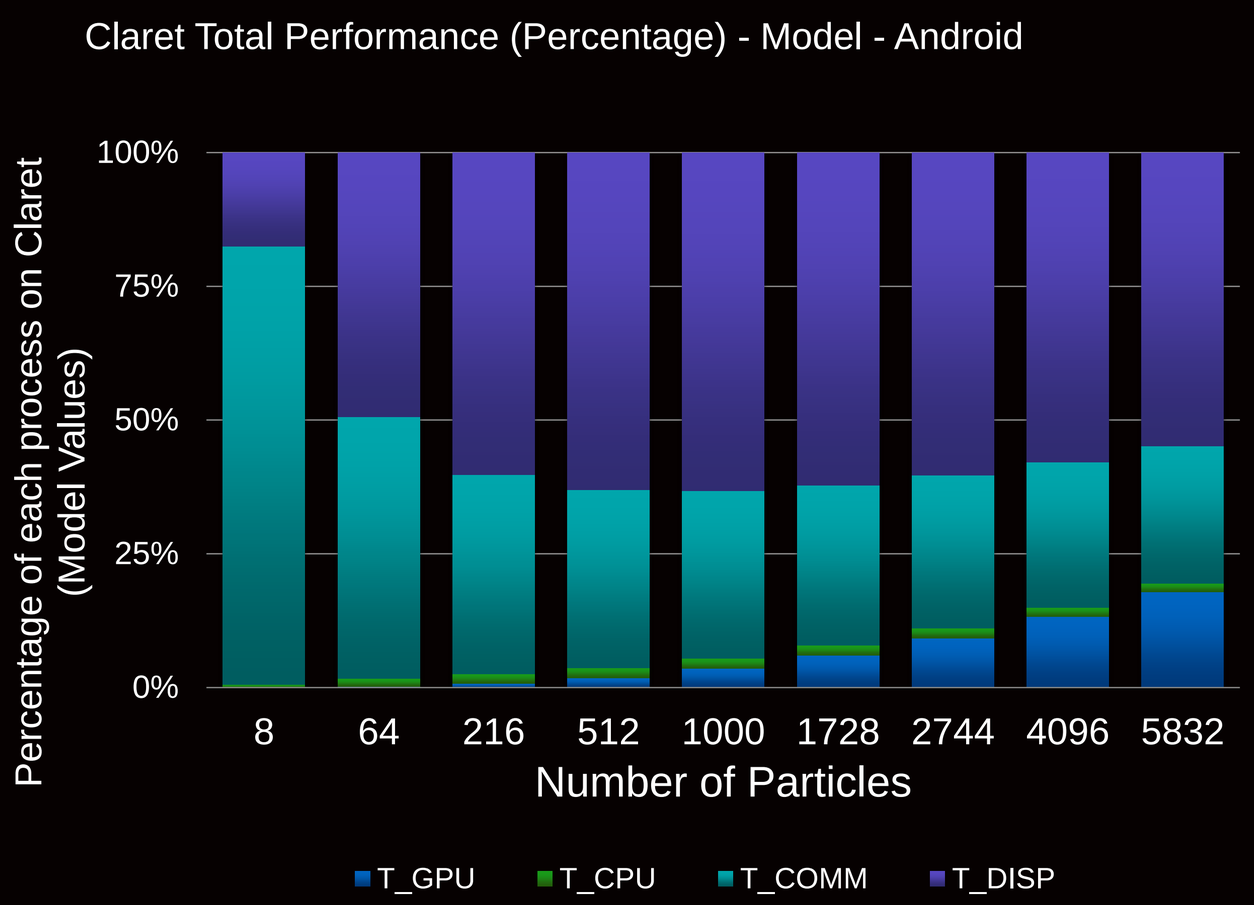
T_DISP: Time for render particles in OpenGL



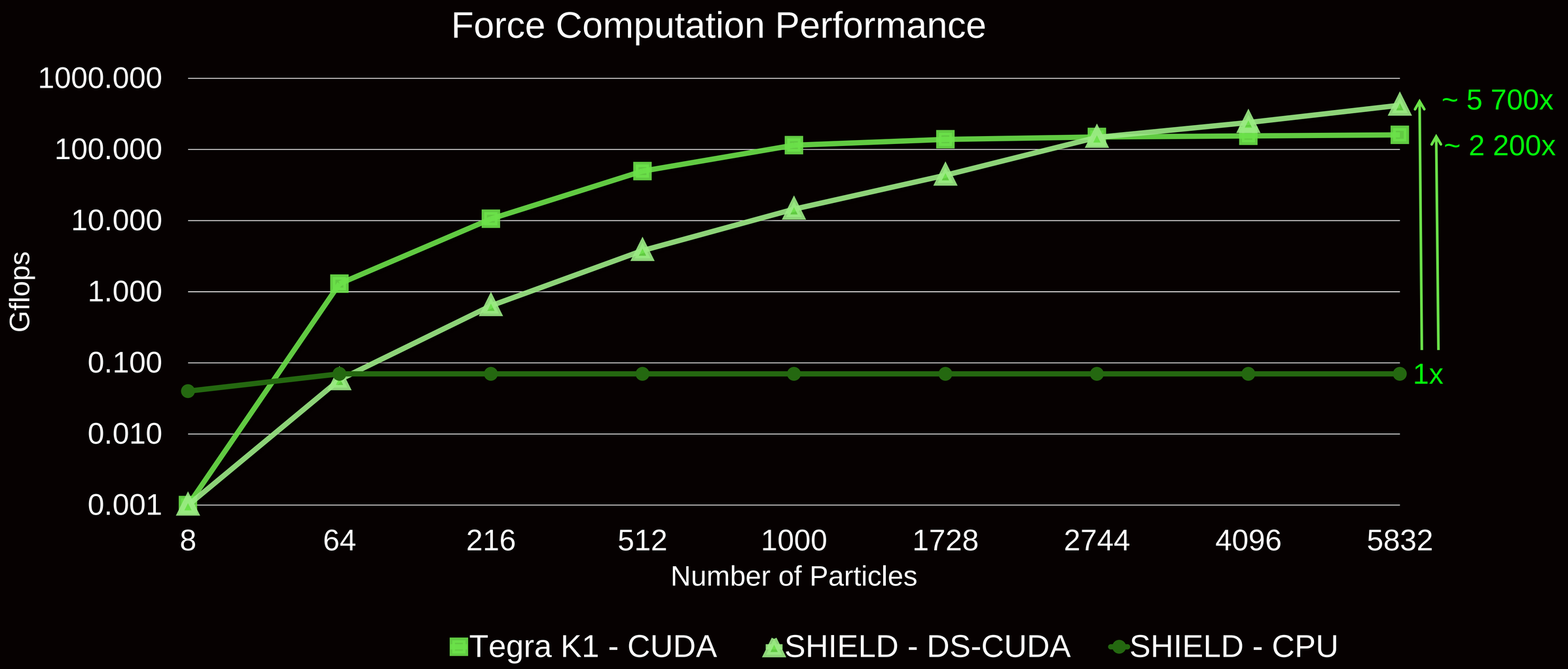
Model of MD simulator for Analysis.

$$T = T\_GPU + T\_CPU + T\_COMM + T\_DISP$$





Tegra K1 vs Tablet SHIELD



Conclusion

- ✓ We were able to run CUDA remotely inside of Android.
- ✓ The usage of HPC frameworks for GPGPU are in development for more than super computers.
- ✓ A molecular dynamics was accelerated inside of the Android Tablet more than 5 000x compared with a CPU implementation.
- ✓ Bottleneck inside of visualization due to:
  - ✓ Many primitives inside of the simulation.
  - ✓ Change for points or textures will be feature work.
- ✓ A study of energy consumption for the tablet is in current progress.

My profile

——————Profile
Name: Martinez Noriega Edgar Josafat (エドガー）
Residence Country: Japan
Current Status: Master Student 2nd Year -HPC
Nationality: Mexican, from Mexico City (Tlaltenco,Tlahuac)
——————Research Interest
High Performance Computing on Mobile Devices
GPU virtualization
Parallel Computing — GPGPU, MPI, MThreading
Molecular Dynamics

Contact:
Email: edgarjosaf@gmail.com
edgarjosaf@uec.ac.jp

LinkedIn: Edgar Josafat Martinez Noriega

Questions???