A Project-based HPC Course for Single-box Computers

Carlos Bederián, Nicolás Wolovick

FaMAF, Universidad Nacional de Córdoba, Argentina

November 14, 2016

EduHPC'16@SC16

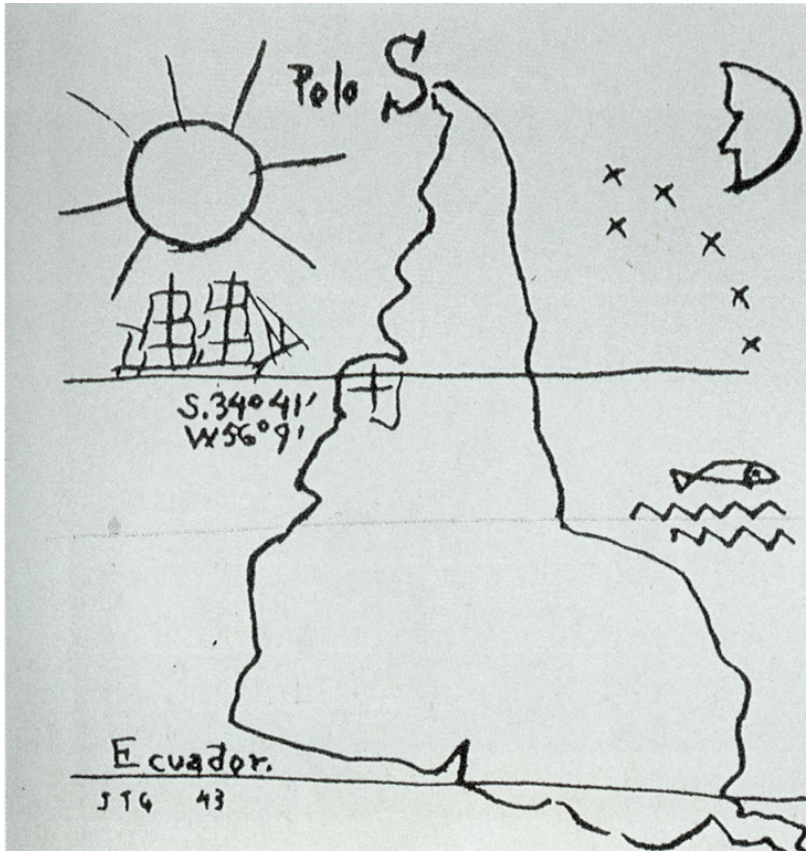
Point of view

How

Challenges

Results

given the fact that we have constructed social space, we know that these points of view, as the word itself suggests, are views taken from a certain point, that is, from a given position within social space. And we know too that there will be different or even antagonistic points of view, since points of view depend on the point from which they are taken, since the vision that every agent has of space depends on his or her position in that space. (Bourdieu 1990c, 130, G: 1992b, 143)



Joaquín Torres García, *América Invertida*, 1943.

Going backwards

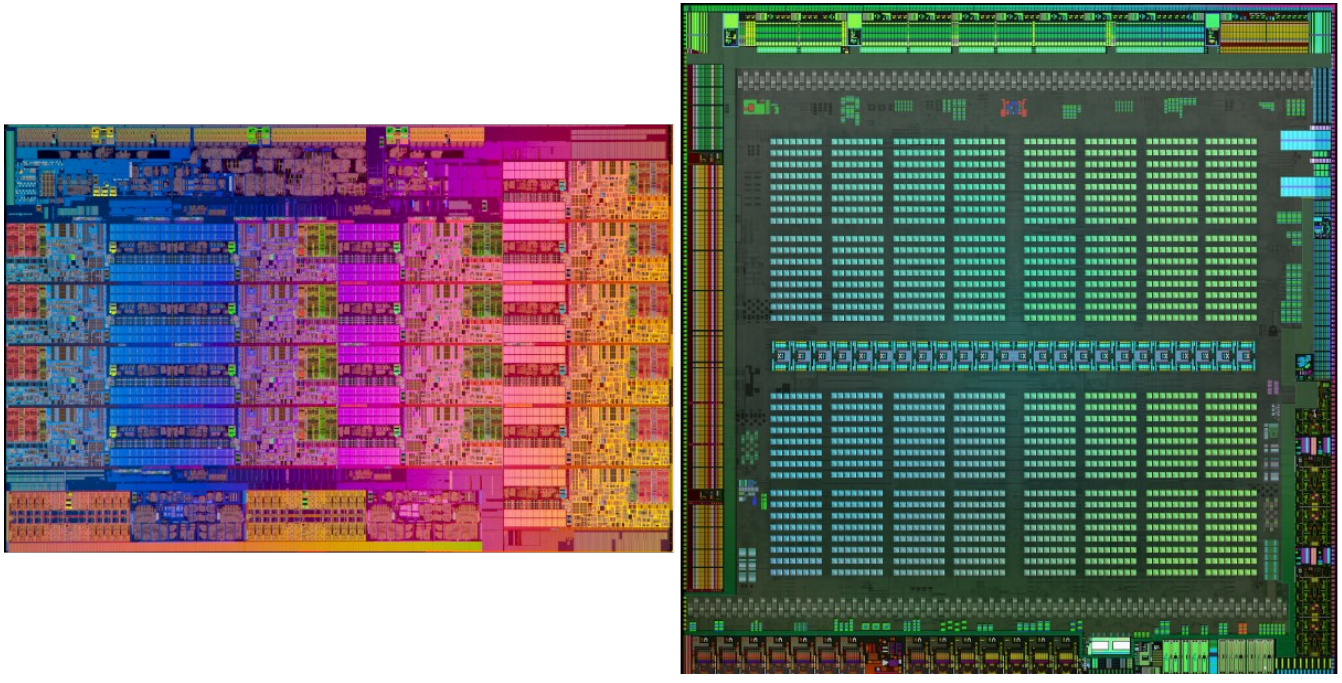
Year	Nickname	Model	R_{peak}	$\frac{R_{peak}}{\min_i\{R^i_{peak}\}}$
1962	Clementina	Ferranti Mercury	5 KFLOPS (sum)	¿?
2000	Clementina 2	Cray Origin 2000	24 GFLOPS	0.681
2001	Deepblue	16×2×PentiumII	25 GFLOPS	0.40
2010	Cristina	70×2×Xeon 5420	5600 GFLOPS	0.24
2010	ISAAC	144×Xeon X3220	5000 GFLOPS	0.178
2014	Mendieta	14×2×Xeon 2680v2	23624 GFLOPS	0.175
2015	TUPAC	58×4×Opteron 6276	48000 GFLOPS	0.265

Argentina, the only G20 member never entering TOP500

Antonio J. Russo, *Computación de Alto Desempeño, Estado del arte en Argentina y en los países del G20*

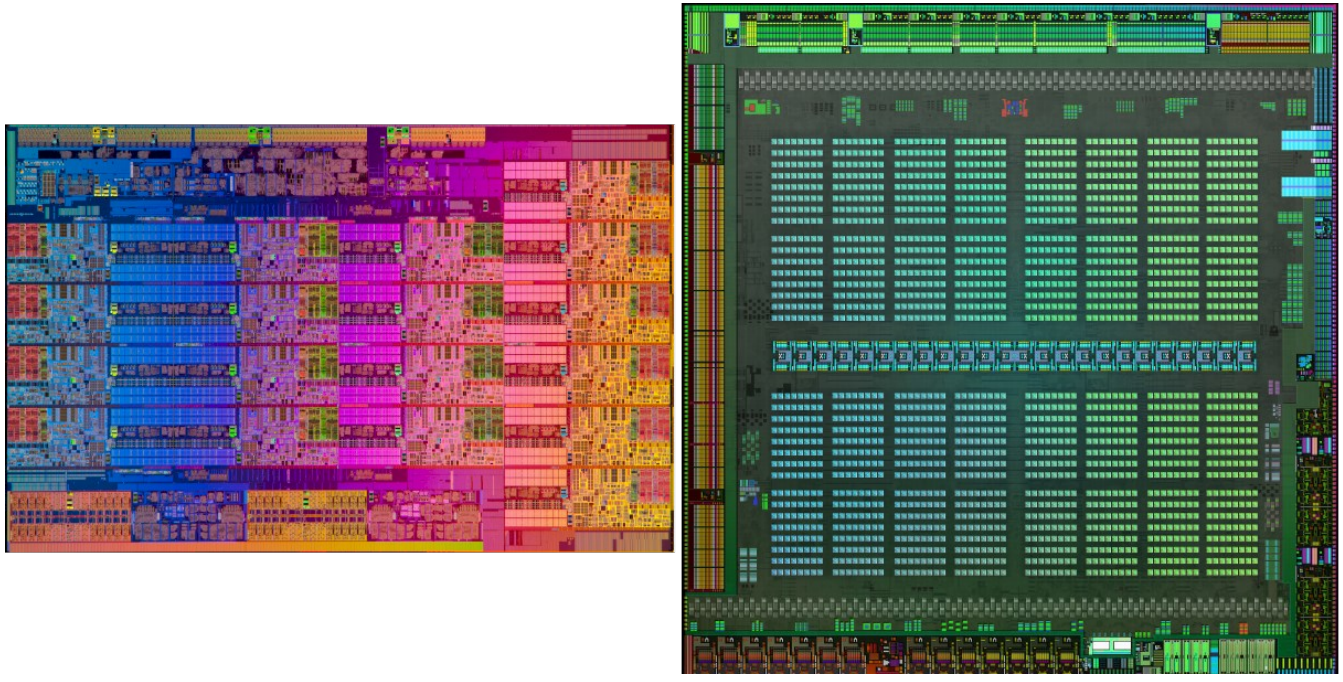
Goal

Sip every transistor out of available architectures

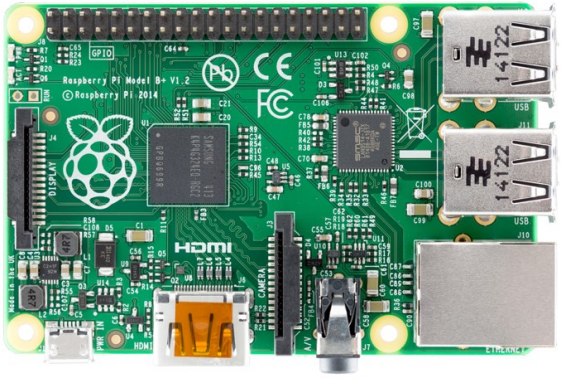


Goal

Sip every transistor out of available architectures



. . . any size



Parallelism

All performance is from parallelism.

Machines are power limited.
(efficiency IS performance)

Machines are communication limited
(locality IS performance)

Bill Dally, *Efficiency and Programmability: Enablers for Exascale*, GTC2013.

Dissecting Parallelism

3 orthogonal dimensions: ILP, DLP, TLP.

- 2 `FMA` ports.
- 8 lanes.
- 12 cores.

192-way parallel CPU architecture.

Dissecting Parallelism

3 orthogonal dimensions: ILP, DLP, TLP.

- 2 `FMA` ports.
- 8 lanes.
- 12 cores.

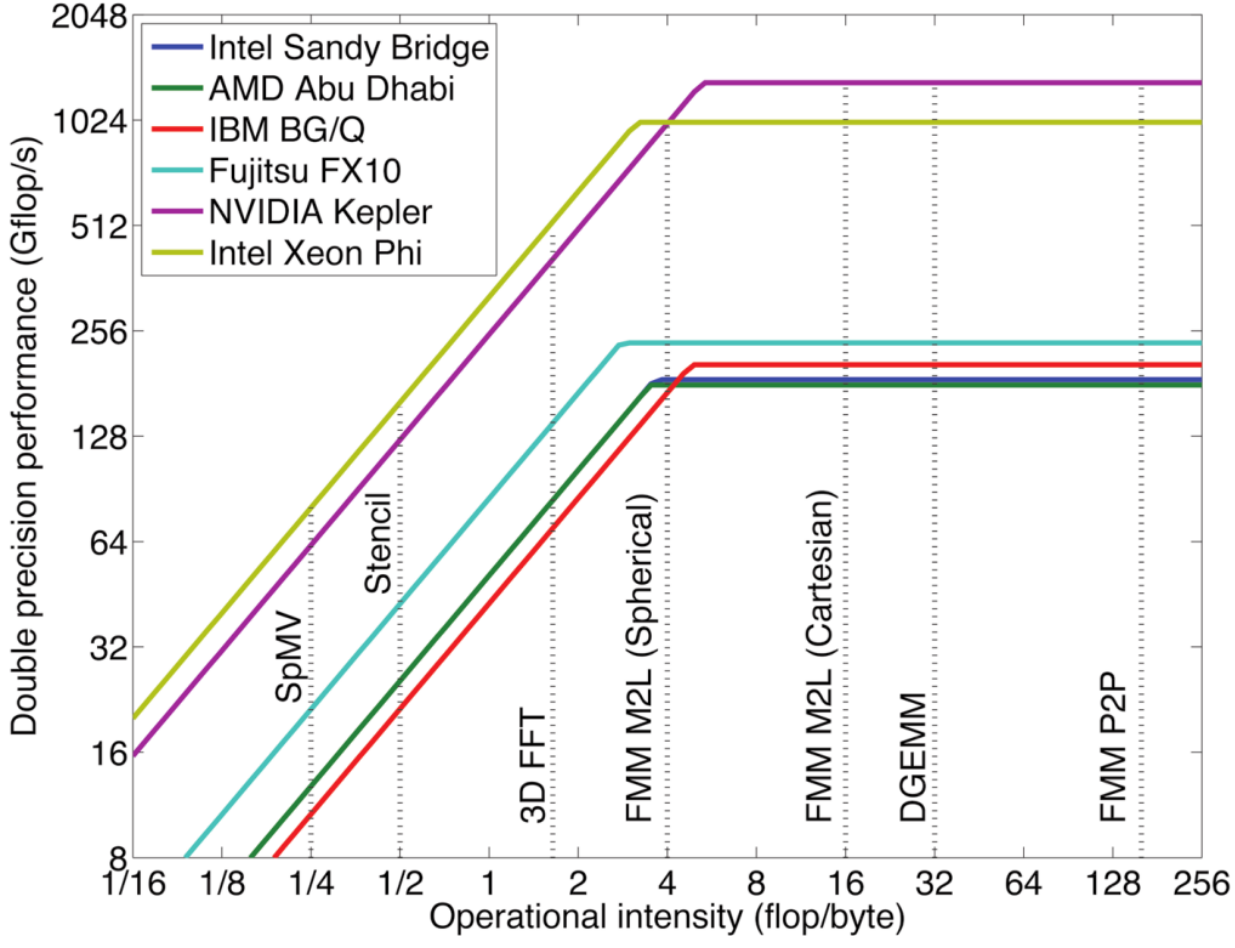
192-way parallel CPU architecture.

GPUs

(wildly) Different implementation of ILP, DLP, TLP.

(somehow) Similar ideas.

Limits



L. A. Barba, R. Yokota, *How Will the Fast Multipole Method Fare in the Exascale Era?*

ILP, DLP, TLP and GPU – Details

ILP

Hard to control, compiler switches, also cover memory hierarchy and `perf`.

DLP

Helping compiler, and hand-made (intrinsics).

TLP

Two levels of locality: global (memory), local (registers). Memory system nightmares (ccNUMA).

GPU

Three levels of locality: shader (global), CTA (shared), warp (registers). Latency hiding, hardware-assisted divergent lanes.

Assembly as **alternative semantics** and
what **transistors execute**.

Project-based

Simple, page-long, yet meaningful numerical simulations.

- A single project throughout the course.
- Apply ILP, DLP, TLP, incrementally in CPU.
- Apply them all in GPU.

Get acquainted with the code.

Understand the problems of each form of parallelism.

Measure normalized speed wrt problem size.

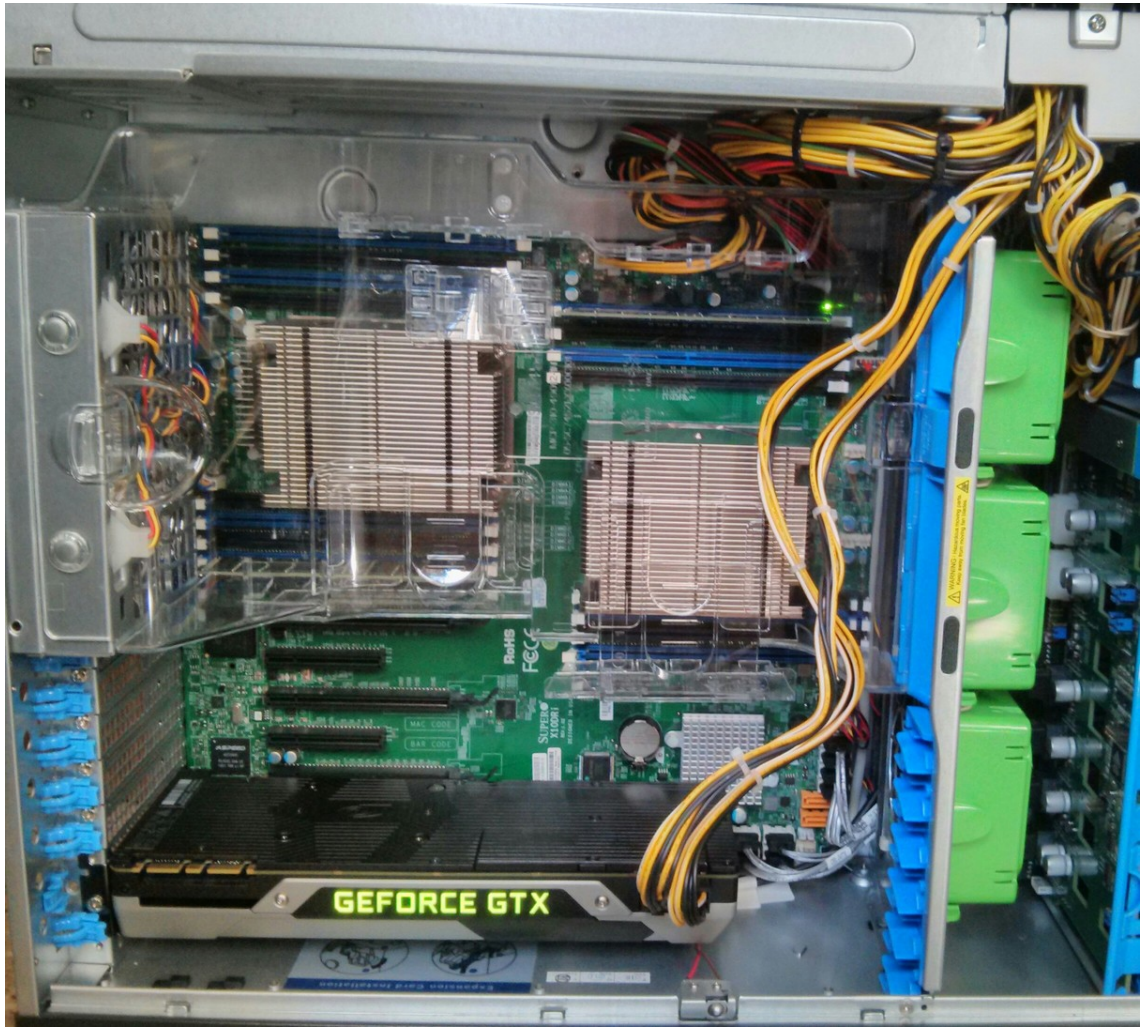
Discuss results after each Lab.

Projects (we trust)

Code	Intensity	Data Dependence	Observations
endoh1	~ 4	Seq/chkboard update	Hard to read source, complex num.
heat	~ 4	Fully parallel update, reduce	Simple
hornschunck	~ 4	Fully parallel	
navierstokes	~ 4	Seq/chkboard update, reduce	Difficult to test for correctness
scan2d	1	Reduction	Hard to beat sequential
spmv	1	Reduction	Random memory access, balancing
tiny_ising	~ 4	Seq/chkboard update, reduce	RNG
tiny_manna	1	Fully parallel	RNG
tiny_mc	N	Fully parallel update, reduce	RNG
tiny_sph	$\sim m$, $m \ll N$	Fully parallel update	Hard-coded, random memory access

Where

Exclusive use & up-to-date hardware:  , 



2×(2×Xeon E5-2620v3, 128 GiB RAM DDR4-2133, SSD 240 GiB, HDD 4 TiB) + 2×GTX980 + GTX Titan X

Bootstrapping

HPCMMM: 2009, 2010

High Performance Computing: Models, Methods and Means(CSC 7600)

ANNOUNCEMENTS

1) Everything due this Thursday!
2) Don't miss the Beyond and Beyond lecture next Tuesday!

PREREQUISITES
Intermediate C/C++ experience
Familiarity with Linux/Unix command-line utilities

LOGISTICS
Location: Room 202, Coates Hall
Timings: Tuesday, Thursday 10:40-12:00

OFFICE HOURS
Tuesday 1:40 -3:00 PM
Thursday 9:00-10:00 AM

MEETING LOCATIONS
Dr.Sterling: Johnston Hall 320
Daniel Kogler: Johnston Hall 318

ARETE Cluster

64 compute nodes and 8 cores per compute node
24 Tb of shared storage
8GB RAM per node
1Gb Ethernet and 10 GB Infiniband interconnect

High Performance Computing

Supercomputing, or more formally "High Performance Computing" (HPC), is the extraordinary new tool of mankind for predicting the future and controlling our world, complementing and augmenting the foundation paradigms of the scientific method of theory and empiricism. In a single life time, the capabilities of supercomputers have grown by more than a factor of a billion; greater than any other technology performance. The participating student will learn the concepts, methods, and means of HPC through a series of hands-on examples, exercises, and assignments to manage, apply, and evaluate the use of these greatest of all computers to real world problems. This course is being conveyed in a multi-media environment for maximum student convenience, accessibility, and interest. The course is being taught in high definition digital video via the internet with Access Grid technology com in human history. HPC is an interdisciplinary field combining digital electronics,computer architecture, system software, programming languages and tools, and algorithms and

Thomas `tron` Sterling.

Loading `initrd`

Computación Paralela: 2012, 2014, 2016

Universidad Nacional de Córdoba
Facultad de Matemática, Astronomía y Física

Computación Paralela 2012

Novedades

- 20120410: charla invitada, martes 15 de Mayo: Daniel Gutson, "Optimizaciones GCC en middle-end (Graphite) y back-end (ARM arch)".
- 20120406: Intel nos dió 25 licencias para uso en clase de Parallel Studio XE y Cluster Studio XE (icc 12.1.3).
- 20120321: Esta página con su contenido inicial.

más

Información General

Docentes.


Carlos Bederián Nicolás Wolovick

Curso optativo de la Lic. en Ciencias de la Computación, FaMAF, Universidad Nacional de Córdoba.

Forma de aprobación: entrega y aprobación de **todos** los laboratorios.

Horarios. **Teóricos**: martes 14 a 16:30, aula 15. **Labs**: jueves 18 a 20, lab 30.

Background diversity

- CS (mostly undergrads)
- Astronomers
- Physicists
- Chemists
- Applied mathematicians
- Engineers

Lots of CS filigree

Tons of theoretical content, but Lab-oriented.
Getting the main ideas is more than enough.

CS-nonCS pairing

Complement skills.

Generating (good) demand

Samples

Luis, mathematician

Dissassembled the code to check it was properly vectorized.

Carolina, physicist

Moved from Matlab to C and she got 10x boost, on top of that she got 5x more by parallelizing.

Cristian, microelectronics

Port a serial code to CUDA and got 36x. It takes 1 minute to compile against hours in the FPGA version.

Johanna, mathematician

Thread-parallelized a R code of her friend obtaining 10x in 12 cores.

Samples – cont'd

Julia and Facundo, CS

Couldn't improve 2.5x in 4 cores for a CFD code, after measuring with `likwid` they realized full memory bandwidth utilization.

Joaquín and Nehuén, CS and electronics

Starting from 113 Kphotons/s, they end up in 13000 Kphotons/s in the same machine, that is 100x in CPU.

Conclusions

One transversal project get acquainted with the problem.

Heterogeneous groups respect each others discipline.

Single-box computer focus and deepen knowledge.

3dim parallelism understand models of parallelism.

GPUs deconstruct CPU assumptions.

Modern hardware let students feel MPP.

Conclusions

One transversal project get acquainted with the problem.

Heterogeneous groups respect each others discipline.

Single-box computer focus and deepen knowledge.

3dim parallelism understand models of parallelism.

GPUs deconstruct CPU assumptions.

Modern hardware let students feel MPP.

Final goal

Generate demand of HPC in the scientific community.
Provide knowledge in HPC technology to CS students.
They already met!

Conclusions

One transversal project get acquainted with the problem.

Heterogeneous groups respect each others discipline.

Single-box computer focus and deepen knowledge.

3dim parallelism understand models of parallelism.

GPUs deconstruct CPU assumptions.

Modern hardware let students feel MPP.

Final goal

Generate demand of HPC in the scientific community.
Provide knowledge in HPC technology to CS students.
They already met!

Eventually we will enter TOP500 ;)

Fin

Fin

Questions?