Towards Advanced Computing Architecture: HPC + AI + Q

Carlos Jaime BARRIOS HERNANDEZ, PhD.
cbarrios@uis.edu.co
About Me (Using AI Chat)

- Carlos Jaime Barrios Hernandez is a computer science professor and researcher in the field of high-performance computing and large-scale architectures. He is a director of the High Performance and Scientific computing Center (SC3UIS) at Universidad Industrial de Santander (UIS). He has co-founded and organized the Latin American Conference of High-Performance Computing (CARLA) and the Supercomputing and Distributed Systems Camping School (SCCAMP). Today, he is the general chair of the Advanced Computing System for Latin America and Caribe (SCALAC) and part of the board of international collaborations in advanced computing, mainly in HPC and Advanced Computing. He has published research papers on advanced computing, new trends in computing, and parallelism. He was a former DJ in France. Follow social networks. Carlos Jaime is a freeride snowboarder.

- Carlos J. Barrios is doctor in informatics at the Université Côte d’Azur, master in applied mathematics and informatics at the Université de Grenoble, both in France, and systems engineer at the Universidad Industrial de Santander in Bucaramanga, Colombia.

(Last week in Chamerousse, Isère, France)
HPC/Advanced Computing Challenges

Infrastructure
- Post Moore Era Architectures
  - Parallel Balancing, I/O, Memory Challenges
- Dark Silico
- Exascale
  - Computer Efficiency (Processing/Energy Consumption)
- Hybrid Platforms (CISC+RISC+Others)
  - TPUs, ARM...
- Data Management / Data Centric
- Advanced Networks
- Fog/Edge
- HPC@Pocket
  - ... Quantum Computing

Platform
- Programmability
  - New Languages and Compilers
- Computing Efficiency
- Data Movement and Processing (In Situ, In Transit, Workflows)
- HPC as a Service
  - Science Gateways, Containers
- Viz as a Service (In Situ)
- Protocols
- IA and Deep Learning Frameworks
- Quantum Computing

Applications
- IA and Deep Learning
- Algorithms Implementation
- Use of Interpreters (as Python)
- Community versions
- Open Algorithms, Open Data
- Utra Scale Applicatons
- Quantum Applications
- and more...
Top Production Applications in Advanced Computing Systems

Source: https://www.g2.com/
Computer Architecture Support (i.e. QC Support)

From https://eca.cs.purdue.edu/index.html

And Sodhi, Balwinder. Quality Attributes on Quantum Computing Platforms.
Main Topics

• Some Computer Architecture Features
• Open Questions (and our contribution)
• From HPC Architecture to Advanced Computing Architecture
• And more Open Questions..
CPU/GPU+TPU Platform

- Good Points
  - Coarse/Fine-Grained Processing
  - Mixed Dense or Sparse Computation (ideal for A.I.)
  - Numerical Methods addressed Large Dependencies (Memory Latency, Memory Bandwidth) and Regularity (ideal for simulations)
  - Memory and FP advantages to simulate states or specific representation (as in the case of QC)

- Bad Points
  - Efficiency (*)
  - Programmability
  - Market Price (Now)
  - (Very) Specific Use
Simulation + Visualization using CPU+GPU

Multi-GPU DSPH Analysis Project Video
N. Gutierrez, S. Gelvez, J. Chacon, I. Gitler and C. Barrios

https://dual.sphysics.org/
Open Question 1: How to exploit better parallelism to support computing and visualization (AI and Simulation)?
Production visualization: “Pure Parallelism”
Production visualization with “pure parallelism”: same problems that processing

Pure parallelism emphasizes I/O and memory
  High Cost (Efficiency, Performance, Energy)
  Difficult to programming and use

Hardware Disruption
  Accelerators (GPUs, ARM, Xeon Phi)
  Specific Issues (i.e. TPU, 3D Memory)
## In Situ Strategies:

<table>
<thead>
<tr>
<th>In Situ Strategy</th>
<th>Description</th>
<th>Negative Aspects</th>
</tr>
</thead>
</table>
| Loosely coupled  | Visualization and analysis run on concurrent resources and access data over network | 1) Data movement costs  
                     |                                                             | 2) Requires separate resources                         |
| Tightly coupled  | Visualization and analysis have direct access to memory of simulation code | 1) Very memory constrained  
                     |                                                             | 2) Large potential impact (performance, crashes)       |
| Hybrid           | Data is reduced in a tightly coupled setting and sent to a concurrent resource | 1) Complex  
                     |                                                             | 2) Shares negative aspects (to a lesser extent) of others |
Loosely Coupled

- I/O layer stages data into secondary memory buffers, possibly on other compute nodes
- Visualization applications access the buffers and obtain data
- Separates visualization processing from simulation processing
- Copies and moves data

Demands Dynamic Memory
Thightly Coupled

- Custom visualization routines are developed specifically for the simulation and are called as subroutines
  - Create best visual representation
  - Optimized for data layout
- Tendency to concentrate on very specific visualization scenarios
- Write once, use once

Demands Dynamic Memory and a large amount of memory capabilities
Hybrid

- Simulation uses data adapter layer to make data suitable for general purpose visualization library
- Rich feature set can be called by the simulation
- Operate directly on the simulation’s data arrays when possible
- Write once, use many times

Demands Dynamic Memory, a large amount of memory capabilities and specific algorithm approach
Analysis and Visualization is run on I/O nodes that receive the full simulation results but write information from analysis or provide run-time visualization.
Our Contribution

**A (new) Algorithm Analytics**
- Performance evaluation for seeds, steps, buffer depth.
- A definition of metrics.
- A new, more detailed evaluation.
- After those, a new algorithm.

**Platforms with In-Situ and In-Transit Strategies**
- Tightly and Hybrid Approach
- Exascale
- Mixing Processing and Visualization Issues

**Applications (Scientific Real Time)**
- GROMACS, NAMD, FlowVR...
- High Available Autonomous Systems
- Specific Libraries and Frameworks
- Ultrascale Software
- Special In-Situ Tools (NVIDIA® VisIt)
- Deep Learning Applications
- Data Movement

---

Sergio Gelvez PhD. Thesis Visualisation Of Vector Fields In Parallel Environments: In-situ Approach Over Heterogeneous Architectures (Advising by K. Garth and C. J. Barrios, Collaborators: B. Raffin (INRIA), J. Hernández (UniAndes) and B. Hernández (NVIDIA))

Source: generated by code provided by VisLab Uni-KL. Rendered in Paraview.
The (Post) Moore Era

After 120 years... The Moore’s Law is Dead

Jack Dongarra
The Cambrian Explosion in Architecture for AI

Convolution Networks
Recurrent Networks
AR Networks
Deep Learning
New Networks

CPU
RAM
GPU
Storage

Satoshi Matsouka Vision

Convolution Networks

RPI
Nano
TX2
Xavier
FPGA

SiP
ASIC
TPU
V100

Deep Learning

Convolution Networks

RPI
Nano
TX2
Xavier
FPGA

SiP
ASIC
TPU
V100

New Networks
Open Question 2: How to exploit Efficiently the Post Moore Architectures?
Virtualization or Containerization?

Evolution of Virtualization

Virtualization

Containers vs Virtualization?
Our Contribution: Performance Impact in Effective Deployment

- Definition of Computing Efficiency:
  - In terms of Energy + “computing element” + processing
- Definition of Post-Moore Era Architectures
  - Parallelism Support + Efficiency + Sustainability?
- Methodology to Analyze and (to predict) the impact of containerization
- Practical Approach to Scheduling Performance Evaluation

Pablo Rojas Thesis « Study of the deployment and execution of applications on post-moore architectures » co-advising with L.A. Steffennel
TPUs: Tensor Processing Units

Tensor Processing Unit (TPU) is an AI accelerator application-specific integrated circuit (ASIC) developed by Google and NVIDIA specifically for neural network machine learning.
DPU Architecture

NVDIA DPU
Open Question 3: How to Achieve Efficiency and Scalability in HPC Architectures that Support AI and Big Data?
Two Approach to Contribute to Deal with the Question:

• Computing Architectural Approach
• Algorithm Approach
What is the problem?

The need to have increasingly efficient computational resources with better performance, among which are greater processing capacity and memory available for the execution of the training of these models.

The deep learning model training algorithm requires a significant amount of memory that often exceeds the capabilities of the GPU and, in some cases, even the memory of the CPU.

New methods for training the model have been created to solve this problem, such as Model Parallelism, Data Parallelism, and Pipeline Parallelism. However, these methods have required increasingly specialized hardware that does not necessarily reduce the memory footprint, but distributes memory requirements across devices such as servers, GPUs, and TPUs.
COMPUTING ARCHITECTURAL APPROACH

BACKPROPAGATION

DATA PARALLELISM

BACKPROPAGATION

Error-backpropagation ($\nabla F(w_i), \nabla^2 F(w_i)$)

Update $w_i \rightarrow w_i + \alpha \nabla F(w_i)$

where, $w$: weight $k$: epoch

Feedback Calculation ($w_j$)

$E = \sum_{i=1}^{N} (Obs_i - Model_i)^2$

where, $N$: number of data set
$Obs$: observations
$Model$: model results

Neuron

Input Layer Hidden Layer Output Layer

No

Yes

End Training

Input Neuron Hidden Neuron Output Neuron

DATA PARALLELISM

GPU 0

Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6

GPU 1

Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6

GPU 2

Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6

GPU 3

Mini-batch

batch 1

Forward Pass

Backward Pass

Compute Gradients $\frac{\partial E}{\partial x_1}$

Updated Model

Gradients

Average Gradients $\frac{\partial E}{\partial w}$

Update Weights $w_{new} = w - \alpha \frac{\partial E}{\partial w}$

Parameter Server

CPU

GPU

Fully connected classification

Feature extraction

Feature maps

Subsampling

Convolution

Output

3

Input

32 x 32

28 x 28

14 x 14

10 x 10

5 x 5
COMPUTING ARCHITECTURAL APPROACH

MODEL PARALLELISM

PIPELINE PARALLELISM
CPU OFFLOADING

Our Contribution:
A New Parallelization Approach in Deep Learning Using CPU/GPU Architectures for Memory Optimization

Thesis of Alejandro Torres co-advising with Yves Denneulin

Is it possible to optimize memory usage in training deep neural network models by distributing or parallelizing the Pipeline between the CPU and the GPU/TPU?

By distributing the Pipeline between the CPU and the GPU/TPU, can better results be obtained in training times while maintaining the accuracy of the model prediction?

By having greater storage capacities in the CPU memory to use it as an active actor, is it possible to increase the size of the input batch and thus improve the efficiency of the training?

Does using the CPU and GPU/TPU simultaneously during training involve more or less energy expenditure when comparing training time Vs. Accuracy Vs. Power Consumption?
Important Aspects:

• Complexity of deep learning models.
• Optimization of the search for hyperparameters in large-scale architectures.
• Population based training
• Generalized DL models
• Evolutionary algorithms in PBT
• Minimizing memory size in the produced model.
• Using AI to bring world-class specialist expertise to everyone, at lower cost.
• Expert care, anywhere.
Our Contribution: Hyperparameters Approach

- Better understanding of PBT-based training mechanisms using distributed computational architectures
- Framework that implements efficient and scalable PBT mechanisms that, through evolutionary algorithms, allows finding generalizable models that minimize memory consumption.
- PBT techniques allow obtaining more optimal generalized models that consume less memory.

Felix Mejia Thesis, Computational efficiency of the implementation of algorithms in Deep Learning applications for health in large-scale architectures in co-advising with M. Riveill and Collaboration with J. A. Garcia.
Quantum Processor Unit Architecture*

*Simplest Approach
Quantum Computing Over HPC

FIGURE 2. Three macroarchitectures for integrating quantum computing with conventional computing. (a) A local machine remotely accesses a QPU through public cloud network. (b) A network of quantum-accelerated nodes communicate through a common interconnect. (c) A network of quantum-accelerated nodes communicate through both conventional and quantum networks.

FIGURE 3. A component diagram representing the microarchitecture of a HPC-QC node with a common interconnect as depicted in Figure 2(b). The diagram shows the major components needed for the operation of a QPU within the HPC node infrastructure. Individual components are grouped into so-called out-of-band and in-band scopes and are placed on the left-hand and right-hand side of the figure, respectively. The QPU, which contains the qubits and is capable of processing quantum information, is depicted at the lower part, whereas classical information processing components are shown in the upper part of the figure. Several hardware (HW) devices control the QPU environment, which has a direct effect on qubit properties and thus the quality of execution of instructions.
Open Question 4: And Quantum Computing?
Our Contribution: Quantum Computing Theory for Quantum Computing Applications

The term **quantum algorithm** is generally used for those algorithms that incorporate some essential feature of **quantum computing**, such as superposition or entanglement. By using this special features, we can speed up significantly the calculation, that is called **quantum parallelism**.

---

G. Diaz PhD. Thesis about Classica Resources Consumption in Quantum Computing Simulators (Co-Advising with L. A. Steffenel)
Final Note: A New Approach of the HPC/HPDA Platforms for Unified Advanced Computing Support (! Or ?)
Why an Advanced Computing Platform Vision (and Not Only HPC)?

(Inspired by the Accelerated/Hybrid Computing World)

**Programming Approaches**
- Libraries: Accuracy and Acceleration
- Directives: Easily Use
- Programming Languages: Maximum Flexibility
- Programming Interpreters

**Development Environment**
- Versions Store
- Developer Hubs, Community Platforms, Pipeline Environments
- IDE
- Linux, Mac and Windows Debugging and Profiling
- Debuggers, Profiling and Performance Visualizers

**(Open) Compiler Tool Chain**
- Linkers, Assembly in Open Source or Corporate Development
- Enables compiling new languages to platforms, and languages to other architectures

**System (HW/MW/SW) Capabilities**
- Post Moore and Non-Von Newman
- Novel Abstractions and Models
- New Computing
- Classical Computing
The Many-Architectures Challenge: How to exploit better Advanced Computing Architectures (for All)?

From Why Quantum Computing is Integral to the Future of HPC’. By William "Whurley" Hurley, CEO of Strangeworks
Conclusion: HPC/Advanced Computing Systems

What is SC3UIS?

R+D+i National Strategic Areas

.com and .org

gov and .co

International R+D+i

.edu

Application Deployment

Scientific Software Development

SCI- IT Management and Support

Strategic Mediation and Training

Research and Innovation
Where is SC3UIS?
SC3UIS at UIS (@UIS) and Guatiguara Technology Park (@PTGuatiguara)

- Founded in 1948 (Following the German /French Polytechnic Model)
- Public State University
- 8 Campus in the Department
  - 4 at Metropolitan Zone of Bucaramanga
  - 4 in Other Regional Cities (Barrancabermeja, Socorro, Malaga, Barbosa)
- 25000 Students (2300 Postgraduate Students)
- 530 Faculty (4 at SC3UIS)
- Support and R+D+I and General Training of SC3UIS

Guatiguara Site was created in 1989 (New Foundation at 2007 as Technology Park)
- 8 Industrial Corporations
- 3 National Labs
- National Core Repository and ANR Site
- 5 Centers
- High Performance Computing Data Center
- GUANE-1 and CHAMAN are here!
- R+D+I and Specialized Training Site of SC3UIS
R+D+i Axes
(@PTGuatiguara)

2017 Important Numbers

4 Patents
5 Spin Off in Incubation Process
(Potentially for 2018 more than 10)
3 Big International Collaborations (more than 5M USD)

2018 New Axes:
Healthcare
New Generation of Automotive Motors
Human and Social Development
About the Use of GUANE-1

Please, follow the simple instructions in
http://wiki.sc3.uis.edu.co/index.php/%C2%BFC%C3%B3mo_Acceder_a_los_Recursos%3F

ssh YourAccount@toctoc.sc3.uis.edu.co
Your password (first time, say yes )

ssh guane

Para mas informacion de como utilizar la plataforma: http://wiki.sc3.uis.edu.co/
Interactive mode:
srun -n xx --pty /bin/bash

List of modules:
  module av
Load Modules:
Recommended:
  devtools/gcc/9.2.0
  devtools/mpi/openmpi/4.1.2
Gracias...

Follow us: @sc3uis