Innovative Applications and Technology Pivots – A Perfect Storm in Computing

Wen-mei Hwu
Professor and Sanders-AMD Chair, ECE, NCSA
University of Illinois at Urbana-Champaign

with

Jinjun Xiong (IBM, C3SR Co-Director), Abdul Dakkak and Carl Pearson

ECE ILLINOIS
Agenda

• Revolutionary paradigm shift in applications
• Technology pivot to heterogeneous computing
• Cognitive computing systems research
A major paradigm shift

- In the 20th Century, we were able to understand, design, and manufacture what we can measure
  - Physical instruments and computing systems allowed us to see farther, capture more, communicate better, ...
A major paradigm shift

- In the 20th Century, we were able to understand, design, and manufacture what we can measure
  - Physical instruments and computing systems allowed us to see farther, capture more, communicate better, understand natural processes, control artificial processes...

- In the 21st Century, we are able to understand, design, and create what we can compute
  - Computational models are allowing us to see even farther, going back and forth in time, learn better, test hypothesis that cannot be verified any other way, ...
# Examples of Paradigm Shift

<table>
<thead>
<tr>
<th><strong>20th Century</strong></th>
<th><strong>21st Century</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Small mask patterns</td>
<td>Optical proximity correction</td>
</tr>
<tr>
<td>Electronic microscope and Crystallography with computational image processing</td>
<td>Computational microscope with initial conditions from Crystallography</td>
</tr>
<tr>
<td>Anatomic imaging with computational image processing</td>
<td>Metabolic imaging sees disease before visible anatomic change</td>
</tr>
<tr>
<td>Optical telescopes</td>
<td>Gravitational wave telescopes</td>
</tr>
<tr>
<td>Teleconference</td>
<td>Tele-emersion – augmented reality</td>
</tr>
<tr>
<td>GPS</td>
<td>Self-driving cars</td>
</tr>
</tbody>
</table>
What is powering the paradigm shift?

• Large clusters (scale out) allow solving realistic problems
  • 1.5 Peta bytes of DRAM in Illinois Blue Waters
  • E.g., 0.5 Å (0.05 nm) grid spacing is needed for accurate molecular dynamics
    • interesting biological systems have dimensions of mm or larger
    • Thousands of nodes are required to hold and update the grid points.

• Fast nodes (scale up) allow solution at realistic time scales
  • Simulation time steps at femtosecond ($10^{-15}$ second) level needed for accuracy
    • Biological processes take milliseconds or longer
    • Current molecular dynamics simulations progress at about one day for each 100 microseconds of the simulated process.
    • Interesting computational experiments take weeks (used to be months)
What types of applications are demanding computing power today?

• First-principle-based models
  • Problems that we know how to solve accurately but choose not to because it would be “too expensive”
  • High-valued applications with approximations that cause inaccuracies and lost opportunities
  • Medicate imaging, earthquake modeling, weather modeling, astrophysics modeling, precision digital manufacturing, combustion modeling, ....

• Applications that we have failed to program
  • Problems that we just don’t know how to solve
  • High-valued applications with no effective computational methods
  • Computer vision, natural language dialogs, stock trading, fraud detection, ...
We know what we want but don’t know how to build it.

Traditional Computer Vision
Experts + Time

Deep Learning Object Detection
DNN + Data + HPC

Deep Learning Achieves “Superhuman” Results

2M training images

Slide courtesy of Steve Oberlin, NVIDIA
Some different modalities of Real-world Data

This seems to be a combinational logic design problem.
## Combinations Logic Specification – Truth Table

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
What if we did not know the truth table?

• Look at enough observation data to construct the rules

  • 000 → 0
  • 011 → 0
  • 100 → 1
  • 110 → 0

• If we have enough observational data to cover all input patterns, we can construct the truth table and derive the logic!
LeNet-5, a convolutional neural network for handwritten digit recognition.

This is a 1024*8 bit input, which will have a truth table of $2^{8196}$ entries.

1M training data is approximately 0%
The adoption of full cognitive business applications has exploded since...
JEOPARDY!
The IBM Challenge
Back in 2011

The cognitive application is built and optimized for the underlying infrastructure manually.

- 90 x IBM Power 750® servers
- 2880 POWER7 cores
- POWER7 3.55 GHz chip
- 500 GB per sec on-chip bandwidth
- 10 Gb Ethernet network
- 15 Terabytes of memory
- 20 Terabytes of disk, clustered
- Can operate at 80 Teraflops
Illinois-IBM C³SR faculties & students (Launched 9/20/2016)
A C3SR App: CELA for Personalized Education

List of available materials

Database of existing science projects

Database of STEM required concepts

Creative Science Project Advisor

Learner’s background Model

Mapping of concepts & projects

Dialog system for Q&A

Video Comprehension

Deep Learning-based

Image Recognition

List of available materials

Materials at hand

Hand Inputs

Camera

Suggested science project experience

Web/text sources for science projects

STEM curriculum, textbooks etc

Deep Learning - based

Experimental Procedure
1. First you need to prepare the 50% glycerol solution, which is made up of half glycerol and half water.
2. Add one cup of glycerol and one cup of water to one of the jars.
3. Tightly secure the lid to the jar and make sure the glycerol is fully diluted, and no goop or lumps remain.
4. Using a permanent marker, label this jar “50% Glycerol”.

Questions (answers) to guide experience

Observing experience

Video

ECE ILLINOIS

ILLINOIS
Extract concept graphs from next generation science standard (http://www.nextgenscience.org/)

• Five blocks of information:
  • Performance Expectations
  • Science and Engineering Practices
  • Disciplinary Core Ideas
  • Crosscutting Concepts
  • Connections

Science and Engineering Practices

Disciplinary Core Ideas

Crosscutting Concepts

The performance expectations above were developed using the following elements from the NGSS document: A Framework for K-12 Science Education:

- Students who demonstrate understanding can:
  - Use materials to design a solution to a human problem by mimicking how plants and/or animals use their external parts to help them survive, grow, and meet their needs.
  - Read texts and use media to determine patterns in behavior of parents and offspring that help offspring survive.

- Five blocks of information:
  - Performance Expectations
  - Science and Engineering Practices
  - Disciplinary Core Ideas
  - Crosscutting Concepts
  - Connections

- Articulation of CCs across grade levels:
  - K-LS1.1, 3-LS1.2, 4-LS1.2, 4-LS1.3, 4-LS1.4, 3-LS1.1, 4-LS1.1

- Connections to other CCs in first grade:
  - Articulate between CCs in first grade:
    - K-LS1.1, 3-LS1.2, 4-LS1.2, 4-LS1.3, 4-LS1.4, 3-LS1.1, 4-LS1.1

- Common Core State Standards Connections:
  - ELA/Literacy:
    - RI.1.1, 2, 3
    - W.1.7

- Mathematics:
  - 1.NBT.B.3, 1.NBT.C.4

- Students who demonstrate understanding can:
  - Use materials to design a solution to a human problem by mimicking how plants and/or animals use their external parts to help them survive, grow, and meet their needs.

- Students who demonstrate understanding can:
  - Read texts and use media to determine patterns in behavior of parents and offspring that help offspring survive.
Paradigm shift for cognitive application development

• Traditional programming approaches failed to deliver cognitive applications for decades

• With the wide adoption of machine learning (deep learning), the core of application development has shifted to model training (including model customization)
  • Experimentation with a large amount of data is on the critical path of application development
  • The nature of functional verification, performance tuning, and debugging is fundamentally different
Cognitive Application Builder (CAB)

A system-level challenge

Workflow description
Innovative AI techniques

High-performance, scalable, robust applications

• CAB: A language, compiler, and runtime for easy development of cognitive applications
  • System-aware to exploit accelerators and efficient communication
  • Introspection for debugging and performance evaluation
  • Workflow optimization and orchestration for system-level performance
  • Decentralized application architecture for scalability, composability, testing, and development
CELA as a driving use case for CAB

- CAB will simplify component connection, workflow description, and iterative development.
• CAB automatically transforms workflows for high-performance execution
C3SR Experimental Heterogeneous Infrastructure

2 x P8 Minsky with NVLink GPUs

4 x P8 Tuleta (S824L)

Courtesy: Jinjun Xiong, IBM
Workload acceleration research at C3SR based on CAB/TANGRAM Software Synthesis

- Focus on impactful cognitive workloads for acceleration
  - Matrix factorization on GPU
  - Long-term Recurrent Convolutional Network acceleration
  - ResNet inference acceleration
  - Neuron Machine Translation acceleration
  - DNN inference acceleration
  - Graph analytic acceleration
- In discussion with other CHN centers to collect performance critical cognitive workloads
- Plan to deliver a set of cognitive benchmarks optimized for OpenPOWER
Matrix factorization: one of key workloads

- **Recommender systems**
  - Predict missing ratings
  - Group similar users/items

- **Natural language processing**
  - Latent semantic model
  - Word embedding as input to DNN

- **Complex network**
  - Link prediction
  - Vertices clustering

- **Deep learning**
  - Model compression
  - Embedding layer

- **Web search**
  - Match query and document

- **Tensor decomposition**
  - In machine learning and HPC applications

In machine learning and HPC applications:

Ratings ($\mathbf{R}$)

- users x items
- $\approx \mathbf{X}\mathbf{\Theta}^T$

- Users
- Items

- Netflix

- Amazon
- Quora
- Apple Music
cuMF acceleration

- cuMF formulation: \( \mathbf{R} \approx \mathbf{X} \cdot (\mathbf{T})^T \) for \( \mathbf{R} \)

  \[
  \mathbf{v} = \sum_{u,v} (r_{uv} - \mathbf{x}_u^T \mathbf{\theta}_v)^2 + \lambda \left( \sum_u \| \mathbf{x}_u \|^2 + \sum_v \| \mathbf{\theta}_v \|^2 \right)
  \]

- Connect cuMF to Spark MLlib via JNI
- cuMF_ALS @4 Maxwell ($2.5/hour)
  \( \approx 10x \) speedup over SparkALS @50 nodes
  \( \approx 1\% \) of SparkALS’s cost ($0.53/hour/node)
- Open source @ [http://github.com/cuMF/](http://github.com/cuMF/)
- Demoed at SC’16 and GTC’16 on Minsky
- Presented to Jen-Hsun Huang on Feb 1, 2017

- cuMF_ALS w/ FP16 on Maxwell and Pascal
- LIBMF: 1 CPU w/ 40 threads
- NOMAD
  - 32 nodes for Netflix and Yahoo
  - 2-10x as fast
Conclusion and Outlook

• Applications have very large appetite for more computing power
  • Both larger scale clusters and faster devices

• Heterogeneity has become the norm for all hardware systems
  • HPC community are currently seeing about 2-3x application speedup
  • Recent positive spiral between deep learning and GPU computing

• Cognitive Computing Systems Research
  • Game changing applications (CELA)
  • Next generation heterogeneous system – democratizing compute and bandwidth (100x)
  • High productivity development with software synthesis (CAB)