Artificial Intelligence at the Edge

Pete Beckman, Nicola Ferrier, Charlie Catlett, Rajesh Sankaran

Co-Director Northwestern University / Argonne Institute for Science and Engineering (NAISE)
Argonne National Laboratory, Northwestern University, University of Chicago
Outline

- Stumbling to the Edge
- A Waggle for Rough Edges
- At the Edge of Chicago
- Science on the Edge
- Cutting Edge Hardware
- Edgy Topics for R&D

Not Today, but find me if you are interested

USA-DOE
JP-MEXT

AMASE
Smart
HPC

Pete Beckman  beckman@anl.gov
Example: SPECIM Camera:
PFD VNIR with 768 bands
(2734 x 1312) x 768 x 2 bytes = 5.1GB image

1 sample every 5 min
Twilight to twilight on June 21 = 1TB

We need a parallel computer with each sensor!
“Out on the edge you see all kinds of things you can't see from the center. [...] Big, undreamed-of things - the people on the edge see them first.”

- Kurt Vonnegut, *Player Piano*
The Edge + Machine Learning
A Revolution

Edge computing and deep learning with feedback for continuous improvement

Powerful Parallel Edge Computing

Semantic Output

Reduced, Compressed data

New inference (program code)

Artificial Intelligence Deep Learning Inference

HPC/Cloud

Deep Learning Training

Pete Beckman beckman@anl.gov
The Edge + Machine Learning
A Revolution

Facility

Actuators
Servos
Dynamic adaptation

Powerful Parallel Edge Computing

Semantic Output
Reduced, Compressed data
New inference (program code)

Artificial Intelligence
Deep Learning Inference

Edge computing and deep learning with feedback for continuous improvement

HPC/Cloud
Deep Learning Training

Pete Beckman beckman@anl.gov
Why Live on the Edge?

- More data than bandwidth
  - Spallation neutron source, light source, HD Cameras, LIDAR, radar, hyperspectral imaging, grid micro-synchrophasors, etc.

- Latency is important
  - Quick local decision & actuation; adaptive sensing & control systems

- Privacy/Security requires short-lived data: process and discard
  - Compromised devices have no sensitive data to be revealed

- Resilience requires distributed processing, analysis, and control
  - Predictable service degradation, autonomy requires local (resilient) decision

- Quiet observation and energy efficiency
  - Vigilant sensors, transmit only essential observations, not big data streams
When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed

When a Computer + Linux is Not Enough…

Challenging Design Contradiction

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed
Smoothing Out The Rough Edges

A Pocket-Sized Controller for Edge Computing

- Borrowed BG/Q control system ideas
- Designed mini “rack controller”
  - Devices can be disconnected
  - Devices can be power cycled
- “Deep Space Probe” design
  - Heart beat signals to each device
  - Alternative boot image / safe mode
  - Current and voltage monitoring
  - Environmental monitoring
- Strict cybersecurity design

The Waggle Manager (WagMan)

Pete Beckman beckman@anl.gov
Waggle: Argonne’s *Edge Computing* Platform

**Bring Parallel Computing to the Edge**

- Supports powerful, parallel computation at the edge
  - Computer vision and deep learning frameworks (Caffé, TensorFlow, OpenCV)
  - Supports edge-optimized & experimental computing
    - ML hardware, GPUs, neuromorphic, FPGAs, etc.

- Open Source, open interfaces

- Integrating advanced sensors easy, with plug-in architecture

- Robust remote system management subsystem

- Manufactured at local electronics company

- ~5 years of development by team at Argonne National Laboratory

Pete Beckman  beckman@anl.gov
The Array of Things Project is Deploying Hundreds of Waggle-based Nodes in Cities

UChicago / National Science Foundation

- 500 nodes will be deployed in Chicago
- Pilot Cities: Denver (Panasonic), Seattle, Portland, Palo Alto, Detroit, Syracuse, Tokyo, Chapel Hill.
- 20+ other cities preparing for pilot projects
- Nodes have 2 cameras, one up, one down
- An instrument to understand urban issues

PI: Charlie Catlett

500 nodes will be deployed in Chicago
Pilot Cities: Denver (Panasonic), Seattle, Portland, Palo Alto, Detroit, Syracuse, Tokyo, Chapel Hill.
20+ other cities preparing for pilot projects
Nodes have 2 cameras, one up, one down
An instrument to understand urban issues

PI: Charlie Catlett
Array of Things Teardown

Edge Computing

Sensor Pod
Current (red) and 60 of 100 additional planned (blue) AoT nodes. Both 1km and 2km buffers are shown, illustrating that even with 200 nodes over 95% of Chicago’s residents will live within 2km of a node and over 75% will live within 1km.

Initial 105 AoT node locations, showing that locations are selected in groups as part of specific science investigations

- **HPC**: Training, Forecast, Optimization, Observation
- **Edge**: Inference, Actuation, Lightweight Learning

GIS map created by A. Laha, Center for Spatial Data Science, University of Chicago
### The Computing Continuum

<table>
<thead>
<tr>
<th>Size</th>
<th>IoT/Edge</th>
<th>Fog</th>
<th>HPC/Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nano</td>
<td>Micro</td>
<td>Milli</td>
</tr>
<tr>
<td>Example</td>
<td>Adafruit Trinket</td>
<td>Particle.io Boron</td>
<td>Array of Things</td>
</tr>
<tr>
<td>Memory</td>
<td>0.5K</td>
<td>256K</td>
<td>8GB</td>
</tr>
<tr>
<td>Network</td>
<td>BLE</td>
<td>WiFi/LTE</td>
<td>WiFi/LTE</td>
</tr>
<tr>
<td>Cost</td>
<td>$5</td>
<td>$30</td>
<td>$600</td>
</tr>
</tbody>
</table>

Count = $10^9$
Size = $10^1$

Count = $10^1$
Size = $10^9$
Transportation

Advanced computer vision to understand pedestrian movement, eventually to predict dangerous interactions with vehicles

Research Credits:
Zeeshan Nadir (Purdue PhD Student @ ANL, 2017)
Nicola Ferrier (ANL Scientist)

Mask r-CNN [He, 2017] trained on COCO dataset

Research Credits:
Yongho Kim, Seongha Park (Purdue PhD Students @ ANL, 2018)
Pete Beckman, Nicola Ferrier (ANL Scientists)

AoT Image from Lake Shore Drive
**Science:** Prototype model of at-grade crossing with impact analysis (interrupt duration and impact; emergency vehicles delayed)

**Objective:** Prioritize among hundreds of at-grade crossings in context of $1B planned investments to improve rail throughput by eliminating key at-grade crossings.

**Science:** Predictive model for multi-modal optimization and control, integrating edge-AI capabilities with traditional transportation data, coupled with HPC models and control systems.

**Deployment:** Integrate transportation measurements from AoT/Waggle (density, flow, vehicle mix, parking) with live traffic data and traffic model around O’Hare International Airport.

**Funding: Illinois DOT**
Partners: Argonne, UChicago, Chicago Metropolitan Agency for Planning, Chicago DOT

**Funding: EERE VTO**
Partners: Argonne, Chicago DOT, Chicago Dept. of Aviation, Chicago Dept. of Innovation and Technology, Arity

Pete Beckman beckman@anl.gov

$3.2m
Hydrology: Flooding

50 consecutive frames to flood water and segment image

Using advanced computer vision to detect surface flooding

Research Credits:
Ethan Trokie (Northwestern Undergrad Student @ ANL, 2017)
Vivien Rivera (Northwestern PhD Student, SCGSR 2018)
Nicola Ferrier (ANL Scientist)
Rajesh Sankaran (ANL Scientist)

Live HPC Flood Modeling and Prediction?

Work with Aaron Packman and William Miller (Northwestern University)
Cristina Negri, Rajesh Sankaran, Nicola Ferrier (Argonne)

Waggle Nodes in Tuley Park, Chicago
Disaster: Flood Fire

Partnership with CSIRO Australia (MOU, visiting postdoc)

- GIS visualisation
- Impact analysis from ensemble wildfire simulations
- Integration with PostGIS database

Live HPC Modeling and Prediction...

• Basis of D61 natural hazard applications:
  • Wildfire and wildfire impact (Spark)
  • Flood and coastal inundation (Swift)
Earth Modeling: Ecosystem Response

Advanced sensors and computer vision to monitor pristine prairie

Research Credits:
Vivien Rivera (Northwestern Univ. PhD Student, 2018)
Aaron Packman, Bill Miller (Northwestern Univ. Professors)
Pete Beckman (ANL Scientist)

Using advanced computer vision to monitor plants

Research Credits:
Renee Zha (Northwestern Undergrad Student @ ANL, 2017)
Zeeshan Nadir (Purdue PhD Student @ ANL, 2017)
Nicola Ferrier (ANL Scientist)

Chicago Botanic Garden Conservation Science Center
Undergrads Caeley and Jordan developed soil moisture sensor now deployed in Chicago
It’s a Bird! It’s a Plane! No… It’s a Drone!

Advanced computer vision and machine learning to identify drones, birds, or fixed-wing aircraft.

Research Credits:
Sean Richardson (USAF visiting ANL)
Adam Szymanski (ANL Scientist)
Benchtop integration of DARPA-funded radiation sensor (Kromek-D3S)
Energy: Power Grid

- Load Forecasting
- Grid Stress
- Air Quality

Precision Micro-Synchrophasors for Distribution Systems: A Summary of Applications

Alexandra von Meier, Member, IEEE, Emma Stewart, Senior Member, IEEE, Alex McEachern, Fellow, IEEE,
Michael Andersen, Member, IEEE, and Laura Mehrmanesh

CRADA Approved!!

Task Order 3 Approval
ASCR Concurrence Received 9/1/2018; OE Concurrence Received 9/11/2018;
SC-32 Concurrence Received 9/13/2018; Using Master CRADA without modification.
Computational Forecasting

Domain setup for HPC forecast

<table>
<thead>
<tr>
<th>Grid/Size</th>
<th>Proc.</th>
<th>Walltime</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 - 32 km²</td>
<td>8</td>
<td>5362s</td>
</tr>
<tr>
<td>130 × 60</td>
<td>16</td>
<td>3041s</td>
</tr>
<tr>
<td>#2 - 6 km²</td>
<td>32</td>
<td>1599s</td>
</tr>
<tr>
<td>126 × 121</td>
<td>64</td>
<td>1033s</td>
</tr>
<tr>
<td>#3 - 2 km²</td>
<td>128</td>
<td>655s</td>
</tr>
<tr>
<td>202 × 232</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DC Forecast

Wind forecasts and measurements

Solar forecasts and measurements

Research Credits:
Emil Constantinescu (ANL Scientist)
Edge to HPC: Live & Historical Data for Forecasting

$$\tilde{y}, \tilde{X}, \tilde{X}, \tilde{y} = m(X_m) + K_{21} (K_{11} + \sigma^2 I)^{-1} (y - m(X))$$

GP & DNNs

Numerical forecast

Historical data

Edge observations used in inference

Validation observations not used in inference

Conceptual Edge solar forecast

Climatology
- Obs for calibration
- NWP
- Observations
- Forecast

Research Credits:
Emil Constantinescu (ANL Scientist)
Flame Spray Pyrolysis

- Use data collected to date to develop ML/DL models
- Relate process parameters to output measures
- Optimize

Research Credits:
N. Ferrier, J.Libera & S. Chaudhuri
Materials Engineering Research Facility, ANL

LLZO transition
Adaptive sampling of the atmosphere

- Atmosphere sensing radars have a wide range of configurations.
- Ideal configuration depends on
  - Atmospheric scene:
    - hurricane, supercell, etc.
  - Phenomena of interest:
    - clouds, tornadoes, birds, bugs
- AI@Edge needed to identify scene
- Automated slicing and dicing to reconstruct spatial structure using machine learning.

Research Credits:
Scott Collis, EVS Division, ANL

New NSF Funding
Facilities: Light Source

Data rates for APS-U will increase several orders of magnitude

Current Experiments:
Times vary: (from seconds to days/weeks)
- Many experimental parameters need to be optimizes
- Data analysis happens after experiment is finalized

Data collection is in the dark
- Parameters are guessed (experience) and then optimized (repeated experiments)

Edge Computing and Experimental Steering
- Improving the science and the efficiency of the experiments
- Real-time data analysis and feedback, data verification, correction, normalization, and configuration parameter optimizations

Research Credits:
Tekin Bicer, ANL

Imaging aluminum foam (dynamic features) sample. Data acquisition and analysis parameters have significant affect on quality of reconstructed feature.

Real-time reconstruction of a shale sample. The scanning pattern and voxel coverage affect the reconstructed quality.
Investments in AI Hardware are Accelerating Change
At BOTH ends of continuum….

Missing: The programming framework for Edge-HPC Science

Google Edge TPU
July 2018

“Edge TPUs are designed to complement our Cloud TPU offering, so you can accelerate ML training in the cloud, then have lightning-fast ML inference at the edge. Your sensors become more than data collectors — they make local, real-time, intelligent decisions.”
Are We Building a “Software Defined Instrument”?

Pete Beckman  beckman@anl.gov
SGP – a multiscale, integrated computational-experimental testbed and a blueprint for other parts of the world
Harnessing The Computing Continuum

Science-driven Problems

e.g.: “Predict urban response to rainfall, trigger intelligent reaction…”

Goal-oriented Annotations

Notional Example:

| trigger {flood_actuation, resident_warning} |
| when {wx_prediction, sewer_model} implies |
| (traffic_capacity < 70%) or (home_flooding > 5%) |

Continuum Abstract Model & Runtime

Existing Resources & Services

Research Focus
Edgy Research: Edge-HPC

- Continually improving Edge-HPC Systems
  - Deep learning + lightweight training + continual improvement
  - Incremental model updates
  - Is Edge really a layer in the model?

- How will the OS/R and system software evolve for Edge-HPC?
  - Scheduling, security, resource management, streaming data

- Programming model & framework for Continuum Computing

- Optimized ML hardware for both Edge & HPC

- Theoretical foundations for failures and correctness of edge/training

- Dynamic resource management and adaptive inference priority
  - AI at the Edge is limited by power and computation – just like HPC

- **Fluid HPC** to support complex and on-demand workflows on future exascale
Questions?

Please drop by for my SC18 Talk

Waggle

http://www.wa8.gl

http://arrayofthings.github.io

Thank You Funding:
- DOE EERE VTO
- Illinois DOT
- Exelon
- NSF
- CSIRO (in kind)
- DARPA (soon)
- ANL LDRD