Computing at the Edge:
Sensors, Learning, and Adaptation

Dan Reed
University of Utah
dan.reed@utah.edu
www.hpcdan.org
Quantitative creates qualitative ...  
... and it changes everything

Viability determined by
- speed, capacity, cost, market scale
... and their ratios

NCSA Cray X-MP (1985)
$8,000,000 and 56 Kbps NSFnet
800 megaflops (peak)
(~$18M in 2018 dollars)

iPhone 8 (2017)
~$700 and LTE wireless
~3000 megaflops

Intel 80386 (1985)
16 MHz @ 2.3 watts
1.5 micron CMOS
Success breeds complacency. Complacency breeds failure. Only the paranoid survive.
Andy Grove, Intel
Our world has changed
Market capitalizations

Billions $ (USD)

Cray, Juniper, ATOS SE, Lenovo, Fujitsu, HPE, Qualcomm, NVIDIA, Broadcom, IBM, Cisco, Intel, Baidu, Tencent, Alibaba, Facebook, Google, Amazon, Microsoft, Apple

“Traditional” computing

Trillion $ (USD) projections

April 2019
Follow the money and the users ...
From petascale to the global race to exascale
Meanwhile, everything gets smart ...

... and the trolley paradox gets real

Oura Ring

Oxford Nanopore Mobile DNA sequencer

IFTTT
The computing continuum: holistic thinking needed

<table>
<thead>
<tr>
<th>Size</th>
<th>Nano</th>
<th>Micro</th>
<th>Milli</th>
<th>Server</th>
<th>Fog</th>
<th>Campus</th>
<th>Facility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>Adafruit Trinket</td>
<td>Particle.io Boron</td>
<td>Array of Things</td>
<td>Linux Box</td>
<td>Co-located Blades</td>
<td>1000-node cluster</td>
<td>Datacenter &amp; Exascale</td>
</tr>
<tr>
<td>Memory</td>
<td>0.5 KB</td>
<td>256 KB</td>
<td>8 GB</td>
<td>32 GB</td>
<td>256 GB</td>
<td>32 TB</td>
<td>16 PB</td>
</tr>
<tr>
<td>Network</td>
<td>BLE</td>
<td>WiFi/LTE</td>
<td>WiFi/LTE</td>
<td>1 GigE</td>
<td>10GigE</td>
<td>40GigE</td>
<td>N*100GigE</td>
</tr>
<tr>
<td>Cost</td>
<td>$5</td>
<td>$30</td>
<td>$600</td>
<td>$3K</td>
<td>$50K</td>
<td>$2M</td>
<td>$1000M</td>
</tr>
</tbody>
</table>

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

Count x Complexity = ~Constant

Stateful vs. Stateless

Source: Beckman, Beck, Dongarra, Ferrier, Reed, and Taylor
Three computing revolutions ...
Deep neural networks at the edge
Most data *never* leaves the device

Amazon Deep Lens
- Integrated camera
- Intel Atom and Ubuntu 16.04
- Intel Gen9 graphics engine
- TensorFlow and Caffe support
- AWS integration (obviously)

Movidius (Intel) Neural Compute Stick
- Custom Vision Processing Unit (VPU)
- TensorFlow and Caffe support
Google Coral ($149)
- Quad-core Cortex-A53 + Cortex-M4F
- *Google Edge TMP ML accelerator*
- 8 GB eMMC
- 1 GB LPDDR4
- GigE, HDMI, USB-C, USB-3
- MicroSD slot

Jetson Nano ($99)
- Quad-core A57
- *128 core nVidia Maxwell*
- 4 GB LPDDR4
- GigE, HDMI, USB 3
- MicroSD slot
## Face recognition on the cheap

<table>
<thead>
<tr>
<th>Hardware Item</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raspberry Pi 2/3 Case (optional)</td>
<td>$9.99</td>
</tr>
<tr>
<td>AZ Delivery Raspberry Pi Camera</td>
<td>$7.99</td>
</tr>
<tr>
<td>Raspberry Pi Model 3</td>
<td>$34.99</td>
</tr>
<tr>
<td>SanDisk 64 GB microSDXC UHS-I card</td>
<td>$11.59</td>
</tr>
<tr>
<td>5V 2.5A Micro USB AC Adapter</td>
<td>$10.99</td>
</tr>
<tr>
<td>Qubo Phone Tripod Camera Stand (optional)</td>
<td>$20.99</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>$96.54</strong></td>
</tr>
</tbody>
</table>
Serious action
Now at the two extremes

The very small (edge/fog computing and sensors)
The very large (clouds,exascale, and big data)

But with real constraints on each

Technical implications
• Fluid end-to-end cyberinfrastructure
• Interdisciplinary data and infrastructure sharing

Cultural implications
• Change management and strategic planning
• Community collaboration
• “Run forever” applications

See the whole board
Deep nets and complex models at the edge
Smart cities and personalized medicine

Connected devices + The Quantified Self

What problem are you solving for?

I want to control my own well being.

Charlie Catlett

Oura Ring

Oxford Nanopore Mobile DNA sequencer

https://arrayofthings.github.io
A framing question ...

How would you use

• Hundreds of ~$50 wireless sensors?
• Streaming environmental data
to understand watersheds and irrigation?

Kang, Kuhl, Bockholt, Rogers, and Reed, “A Cloud-Based Scientific Gateway for
Internet of Things Data Analytics,” PEARC18
What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.

Herbert Simon
Science instrument continuum

BIG and FEW

Multidisciplinary Data Fusion & Modeling

SMALL and MANY

Vehicles

Array of Things

Health

ATLAS

LSST

SKA

NEON

INTERNET of THINGS
Building fluid capabilities: AND not OR

- LHC, LSST, SKA (Big Instrument) and HPC
- Smart City/Environmental Monitoring (Edge)
- Distributed Workflow (Hybrid)

Mapping
- What
- Where
- When

Subject to
- Speed
- Capacity
- Latency
- Resilience
- Security

Edge/Edge  Fog  HPC/Cloud/Instrument

Stateless

Stateful

AWS Greengrass
Harnessing the computing continuum
Intentional specification

Science-driven Problems

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

Notional Example:
trigger \{flood\_actuation, resident\_warning\}
when \{wx\_prediction, sewer\_model\} implies
(traffic\_capacity < 70%) or (home\_flooding > 5%)

Existing Resources & Services

Source: Beckman, Beck, Dongarra, Ferrier, Reed, and Taylor
Building fluid capabilities

Edge
Latency
Bandwidth
Energy

Edge
Storage
Knowledge
Context

Edge

Latency
Bandwidth
Energy

Edge

NFC
5G
Z-Wave
LoRa

4G
LTE
ZigBee

Wi-Fi

4G
LTE
ZigBee

Wi-Fi
Chip feature sizes and Dennard scaling
No exponential is forever (except in the textbooks)

Why us? Why now? (channeling Jim Larus)

Why us?
• Someone needs to think about higher level models

Why now?
• End-to-end, “run forever” services are the future
• Architectural specialization puts a premium on portability
  • Minimizing data movement at many levels
  • Maximizing operations/joule

Why fusion?
• Integration will enable new capabilities
• It’s more than workflows, containers, and libraries
Not yet fully realized ... my 2002 prediction
The big questions don’t change ...

... but the approaches and answers do