IoT Research

Prof. Tajana Šimunić Rosing
Dept. of Computer Science

System Energy Efficiency Lab

seelab.ucsd.edu
The changing nature of computing

- Impossible to manage using centralized control -> distributed resource management and control are key

---

50 Billion connected devices

- More connected devices than people

- Cisco 2011
High Performance Wireless Research and Educational Network – HPWREN

HPWREN wireless connectivity covers 20k sq. mile area with numerous sensors

Motion detect cameras

Wildfire tracking cams
Cal Fire Dept.

Environmental sensors & cams

Acoustic sensors
Wolf howls at the CA Wolf Center

HW Braun, F. Vernon, T. S. Rosing, UCSD
Mount Laguna Sensors

- 3D ultrasonic anemometer
- anemometer
- solar radiation
- tipping rainbucket
- temperature
- relative humidity
- barometric pressure
- data logger
- Pan-tilt-zoom camera
- support equipment
- 3D ultrasonic anemometer
- solar radiation
- tipping rainbucket
- temperature
- relative humidity
- barometric pressure
- data logger
- Pan-tilt-zoom camera
- support equipment
Wind Gusts on Mt. Laguna
U.S. Navy Deep Submergence Unit

Santa Margarita Ecological Reserve

75 Cluster Heads connected via WLAN
Volcan Fire HPWREN connection, September 2005
Goal: seamless analysis and reuse of context

- Applications process raw input data → output actions
  - Avoids processing redundancy & monolithic applications
- Provide ontological information
  - Ontology provides range, discretization, type of input data

Our approach: Replace black-box applications with multi-input single-output functional units – context engines

- Composable units allow data reuse & parallelization
- General purpose statistical learning to generate output based on ontological metadata:
  - Classification: Decision Trees, SVM
  - Regression: KNN, SVR, Linear Regression, TESLA
  - Clustering: K-means, D-Stream II
  - Anomaly Detection
- Simple API facilitates development; includes full RSA
- Runs on small devices (Arduino, RPi3), and on larger systems (laptops, servers)
- Easily scalable and capable of adapting to change
Machine Learning over IoT Network Hierarchy

Motivation:
- Leverage computation capabilities throughout the IoT hierarchy
  - Reduces communication costs -> increases battery lifetime
  - Cloud only has access to derived features -> accuracy

Neural Networks fit perfectly!
- Powerful modeling technique
- Can solve many ML problems
- Model structure fits well into a hierarchy

Towards a Solution:
- Optimal compression through local NNs
- Trading width for depth of NN
  - Formulate bounds on accuracy-communication trade off
- On-line adaptation

Jointly with Prof. Arun Kumar
Context Engine Example: Proactive Fire Tracking in San Diego County

Current state of the art: Stationary sensors & cameras are placed in high risk areas. In 2003 & 2007 San Diego fires, the flames already spread beyond our ability to control them as they started away from the sensors.

Our goal: Develop a proactive infrastructure that leverages both stationary & drone based sensors to sniff out fires & provide adaptive wireless connectivity to the troops on the ground.
Modeling User Behavior for Smart Grid Control at Large Scale

Automated framework to model human behavior for design of context-aware distributed control in Smart Grid applications across various geographies and for all key demographics of interest.

- **Workload modeling**
  - Create footprint profiles based on human activities

- **Quality of service modeling**
  - Find user flexibility regions
  - Classify the CDFs with respect to entropy
  - Flexibility ranges based on timing

- **Scheduling**
  - Application to neighborhood:
    - Consider user flexibility
    - Reduce overall cost
    - Minimize the peak power

**Tradeoffs:** cost vs. peak power:
- 16% cost ↓ -- 22% peak power ↑
- 12% cost ↓ -- 11% peak power ↓
- 1% cost ↑ -- 5% peak power ↓

**ATUS:** Activity modeling – over 10000 people/year from entire USA

**RECS:** Appliance use modeling – over 110M participants from entire USA
NSF CitiSense: Wearable and Environmental Sensors

- Sensors and phones given to commuters using various transportation means throughout the greater San Diego area
- Results found “urban valleys” where buildings trapped pollution
- Major routes reported contrastingly high/low AQI for the same location
- Ease of deployment and in-network adaptation is key: learn from context!
Proactive Health at Scale: Population & Individual Context for Personal Feedback re. Health Risks

Enables citizens to take personal sensor readings and view them in the larger context of population-level knowledge for personalized health insights.

**Population Level Analysis**
- Per-county demographics, habits & environment data

**Personal Sensing**
- Current environmental exposure (e.g., exercise, daily pollutant exposure & movement etc.).
- Seamless and adaptive integration of new components

**Proactive Feedback**
- Personalized suggestions based on individual data, aggregated local information & population level models

"Let's Get Healthy California" Task Force: Asthma ER admittance per 10,000

Correlation with ER admittances

Carbon monoxide:
- black – reference EPA monitor
- colors – fitted sensor models

In collaboration with UCSD School of Medicine
For the first time in history, we will longitudinally study the impact that daily habits, the environment, genetics, and the microbiome have on older adult cognition. **Goal:** We will model the subtle changes of aging, and deploy personalized health interventions via cognitive robots to support independent living.
Hidden Costs of IoT

• Initial costs:
  – Design, HW sourcing and manufacturing, implementation

• Operational Expenses (OpEx) are recurring costs:

<table>
<thead>
<tr>
<th>Network communication</th>
<th>Administrative labor</th>
<th>Technical support</th>
<th>Time to market</th>
</tr>
</thead>
<tbody>
<tr>
<td>33-50 percent</td>
<td>20-50 percent</td>
<td>10-33 percent</td>
<td></td>
</tr>
<tr>
<td>Monthly/device</td>
<td>Every deployed</td>
<td>10% of deployed</td>
<td></td>
</tr>
<tr>
<td>subscription fee</td>
<td>device requires</td>
<td>devices require</td>
<td></td>
</tr>
<tr>
<td>Monthly/device</td>
<td>at least 15</td>
<td>support</td>
<td></td>
</tr>
<tr>
<td>overage fee</td>
<td>interactions/year</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Time to provision</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>devices and services</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing devices</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>and services before</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>deployment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MTTR: 25 minutes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T1 MTTR: 25 minutes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T2 MTTR: 3 to 5 hours</td>
</tr>
</tbody>
</table>

Source: Jasper, Cisco
Average Annual Administrative Costs for 100,000 Devices

- Data analysis of more 3500 companies worldwide
- Assumptions:
  - Average salary for professional operations administrator in the USA is $48/hr
  - Each interaction requires 5 minutes of labor
  - Number of interactions varies by industry:

<table>
<thead>
<tr>
<th>Industry</th>
<th>Touches</th>
<th>Cost (Before Platform)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security &amp; home automation</td>
<td>13</td>
<td>$5.2M</td>
</tr>
<tr>
<td>Point-of-sales company</td>
<td>15</td>
<td>$6.0M</td>
</tr>
<tr>
<td>Industrial equipment</td>
<td>5</td>
<td>$2.0M</td>
</tr>
<tr>
<td>Transportation &amp; logistics</td>
<td>7</td>
<td>$2.8M</td>
</tr>
</tbody>
</table>
Average Annual Technical Support Costs for 100,000 Devices

- Data analysis of more 3500 companies worldwide
- Assumptions:
  - Average tier 1 engineer salary is $20/hr, and call time is 25min
  - Average tier 2 engineer salary is $40/hr, and the mean time to resolution is 4hrs
  - 20% of support has to be escalated from Tier 1 to Tier 2
  - 10-30% of devices require technical support

---

**Without platform:**

- **Security & home automation:**
  - 30% require technical support
  - Without platform: $1.45M
    - $250K for T1
    - $1.2M for T2

- **Point-of-sales company:**
  - 20% require technical support
  - Without platform: $967K
    - $167K for T1
    - $800K for T2

- **Industrial equipment:**
  - 25% require technical support
  - Without platform: $1.2M
    - $208K for T1
    - $1M for T2

- **Transportation & logistics:**
  - 10% require technical support
  - Without platform: $483K
    - $83K for T1
    - $400K for T2
Sources of failure

- User errors
- Installation errors
- Communication problems
- Power issues (battery lifetime)
- Hardware failures
  - Soft errors: transient
  - Hard errors: permanent
Power, Thermals, Reliability & Variability

Interrelated problems!

- Multiple power hungry subsystem
- Intense application requirements
- Limited battery capacity

Workload User Ambient temperature

Power dissipation increases temperature
Big issue with no active cooling (fans)

Degradation mechanisms

Temperature

Reliability

Utilization and temperature stress reduces device lifetime

Variability

- Performance
- Power
- Degradation Rate

Performance-related metrics

- Technology is rapidly scaling
- Accuracy of fabrication process is limited
A comprehensive management framework for power, temperature, reliability, & variability management in individual devices

User Experience Requirements

Comprehensive Management Framework

- Variability
- Reliability
- Temperature
- App perf
- Battery
- Power

Online Reliability Emulation

Online Variability Emulation

System feedback

Control decisions

Target system

LTC: Long Term Controller
STC: Short Term Controller
Networks of IoT Devices

Internet of Things:
- Energy and reliability are important issues -> maintenance costs
- Challenges: heterogeneity, distributed nature, large scale

Research directions:
- Framework for energy / reliability management of the IoT infrastructure
- Control of multiple heterogeneous variables
- Automatic recognition of user experience requirements

![Diagram of a network with layers and nodes labeled with reliability, power, QoS, utilization, and control decisions. The diagram includes connections and streams labeled with 'Utilization log stream' and 'Control decisions stream'.]
Acceleration of IoT workloads: CPUs, GPUs, DSPs, PIM

Approximate Computing HW Supports

Enhanced CPU
- Associative MEM

Enhanced GPU
- ReCAM
- MASC

Enhanced DSP

Ap Storage

NDC* Accelerator

Original Computing Units

Main Memory

New NVM-based Components

Relaxed MEM

Streaming App

Application Framework

System Libraries
Library of PIM Accelerators

Clustering
- Kmeans
- HD Clustering

Classification
- Hyperdimensional Classifiers
- Adaboost
- DNN, CNN
- Decision Tree
- kNN

Supporting both Training and Testing

Graph Processing
- Graph Processing

Database
- Query Processing
Example: NNgine, kNN Accelerator

- **NNgine**: Nearest neighbor search accelerator
- **Applications**: pattern recognition, statistical classification, biology, computer vision, databases, coding theory, computational geometry
- Performing the search operation inside the magnetic memory next to DRAM

<table>
<thead>
<tr>
<th>NNgine Energy Improvement vs. GPU</th>
<th>5590x</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNgine Speed up vs. GPU</td>
<td>510x</td>
</tr>
</tbody>
</table>

*NNAM: Associative memory with capability of finding k-nearest neighbor
Hyperdimensional computing:

- Represent data using bits encoded at large dimensionality (>10,000); such randomly encoded vectors are orthogonal
- Purely statistical, thrives on randomness
- Supports full algebra

Superb properties:

- General and scalable model of computing
- One-shot learning
- Robust against most failure mechanisms and noise
Example: DNNs using Hyperdimensional Computing

In memory acceleration as compared to ISSAC [ISCA'16]:

**Analog HD:**
- $1417 \times$ higher energy efficiency & $135 \times$ speedup
Overview of SEE Lab Projects

- **Focus is on energy efficiency across scale**
  - Chip multi-processors:
    - Ultra-fast power gating of CMPs
    - Energy management of heterogeneous CMPs
  - Smaller scale systems
    - Heterogeneous memory integration and management
    - Energy management in mobile phones
    - Power/thermal/cooling management
  - Large scale systems
    - Energy management in data centers
    - Green energy modeling and management
    - Heterogeneous wireless sensor nets (CitiSense, context awareness)
    - Distributed sensing and control for SmartCities with focus on urban sensing and smart grid
    - Healthy aging in place