Sensor Data Engineering
Some Lessons Learned

Philippe Bonnet, IT U.Copenhagen - phbo@itu.dk

Joint work with: Praveen Seshadri, Martin Leopold, Marcus Chang, Klaus Madsen, M.Bjørling, Luc Bouganim, Bjørn Jonsson.

Work funded by danish strategic research council, danish research agency special program on IT, EU FP6 CA/NoE.
Sensor Data essentially used to model physical systems. Those models dictate how sensor data should be collected and managed.
Sensor Data Engineering

Engineering techniques and methodologies for the design, development and assessment of data acquisition and sensor data management systems.
Sensor Data Lifecycle

- **Instruments**
  - High-bandwidth – not all captured data part of raw measurements (particle physics, vulcano monitoring)
  - Low-bandwidth – all data are captured (ecoligical datasets)
- **Raw measurements are primary data, data products are secondary data**
  - e.g., NASA CODMAC
- **Instrument Control Loop**
  - Parameters adjustment based on model knowledge
- **Archiving/Curation of cold data.**
Example Data Products

NASA CODMAC:

**Level-0**: raw engineering measurements (full resolution)

**Level-1**: raw measurements time-referenced and with meta-data computed, appended but not applied

**Level-2**: Derived geophysical variables at the same resolution

**Level-3**: Variables mapped on uniform space-time grid scales

**Level-4**: Model output or results from analyses of lower level data
Outline

Sensor Data Engineering

Sensor Network-Based Instrumentation

Hogthrob: Characterizing Mote Performance

Mana: Autonomous Data Acquisition

Where do we go from here?
Rationale #1

Sampling at unprecedented scale and resolution

Astronomy showing the potential of inexpensive sensor + computing

Graph courtesy of A.Szalay
The MEMS\textsuperscript{(*)} Revolution

\textsuperscript{(*)} Micro Electro-Mechanical Systems

Evolution in time of the form factor of accelerometers.

Form factor, cost and energy consumption are highly correlated.

**Reality Check:** Exponential evolution in the size of sensor driven by automotive, consumer goods and medical devices industries. True for Accelerometers, gyroscopes, imagers, microphones. Not true for optical, Chemical, biological sensors.
Rationale #2

Sensornet-Based Instrumentation (a) introduces automatic observations in remote places where only manual observation have been possible so far, and (b) transforms stand-alone devices into a networked system that is monitored and controlled for better performance.
Mote-Based Instrumentation

System challenges: (1) Prototyping tomorrow's platforms 
(2) Exploring the boundaries of today's platforms
Design Space

1. Choose HW components
2. Choose SW components (incl. OS)
3. Implement new components
Design Space

Sensor network regime

- Connectivity vs. Duty cycling
  - Low-Power Listening
- Topology control
  - Gateway + cluster of 20-50 motes
- Network abstraction
  - IP (6LowPan)
- Routing: open issue
Software Development (Data Acq.)

- Programming abstractions
  - **Concurrency**: thread vs. events, time representation
  - **Distribution**: primitives (reliable dissemination, collection tree), macroprogramming

- **Testing**: controllability and observability are hard/expensive to enforce when instrumenting a physical system. Verifying calibration, testing compliance after deployment are open issues.

- **Health management**: making sure data acquisition functions and performs as required once deployed is an open issue.

- **Design Principles**: Equivalent of Rules of thumbs in data engineering [Gray], Hints of system design [Lampson]
Data Acq./Mgt Paradigms

Sensor specific

- Prob. DB [MayBMS, Trio]
- Autonomous Data Acquisition [iDare]
- StreamDB [Aurora, Stream, ...]

Model-based query processing [BBQ, COLR]

Delay-tolerant query processing [IceDB]

Gateway-based Systems [SwissQM, Tenet]

RDBMS

In-network query processing [Cougar, TAG]

Separated

Generic

Integrated
Software Development (Data Mgt.)

- Interesting times!
  - Tape is dead, disk is tape, flash is disk [Gray]
  - RAM is disk, L2 is RAM on multicore [Ailamaki]
  - Drop disks and forget about buffer manager, locking and logging [Stonebraker]

- Sensor data are uncertain spatial time series
  - Impacts data model, query language [Widom]
  - Impacts storage structures [Madden]

- Loading from Data Acq to Data Mgt is key [Howe]

- Supporting the complete sensor data lifecycle is an open issue.
Outline

Sensor Data Engineering

Sensor Networks

Somes Lessons Learned: Hogthrob, Mana.

Where do we go from here?
Hogthrob

- Infrastructure for **sow** monitoring
- Today, RFID-based ear-tags to monitor food-intake
- Tomorrow:
  - Life-cycle monitoring: heat, farrowing, localization, injury detection.

**Consortium:** DTU, KVL, DIKU, National Committee for pig production
What if we were to use a SoC mote?

- **Rationale**: Meets form factor requirement
  - Promises to meet cost, lifetime requirements
  - Feasible: Spec at UC Berkeley [J.Hill 2002]

- **Problem**: How to explore the design space?
  - Should we build a mote, should we pick an existing mote? Can we use a CC2430 based Nano mote instead of the MSP based Micro?
  - Vector-based approach to characterize mote performance [Leopold et al. 2008]
Hogthrob WSN

Functionality: Data Acquisition so that veterinarians can model the behaviour of group-housed sows.
Deployment#1  (Feb 2005)

- **Goal:**
  - 3 weeks
  - 5 sows
  - Collect time series
    - 2D and 3D accelerometers
    - 4 Hz sampling ~200 MiB / mote
  - Video as ground truth
    - BW Surveillance camera (AVC301a)
    - 8KiB per PNG frame
    - 32 KiB/sec
    - Total: 80 GiB per camera
Deployment #1

• Approach
  - Using BTNodes V2 running TinyOS
    • Bluetooth radio
    • Internal flash (~100KiB)
    • Aggressive duty cycling: sample and transmit when flash is full
    • 4 NiMH cells (1.25V, 2400mA)
  - Logical time stamp on motes - physical time stamp on gateway
Deployment #1

• Results
  - Lifetime requirement met
  - Disappointing yield
  - Painful post-processing
    • Node reboots
    • Time difference across the two gateways

<table>
<thead>
<tr>
<th>Node</th>
<th>First Measurement</th>
<th>Last Measurement</th>
<th>Days</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>4D08</td>
<td>2005/02/28 14:05</td>
<td>2005/03/13 20:31</td>
<td>14</td>
<td>44.607%</td>
</tr>
<tr>
<td>4D09</td>
<td>2005/02/28 09:59</td>
<td>2005/03/19 22:47</td>
<td>20</td>
<td>71.756%</td>
</tr>
<tr>
<td>4D0B</td>
<td>2005/02/28 10:06</td>
<td>2005/03/19 22:48</td>
<td>20</td>
<td>64.567%</td>
</tr>
<tr>
<td>4EBC</td>
<td>2005/02/28 10:28</td>
<td>2005/03/19 21:06</td>
<td>20</td>
<td>56.139%</td>
</tr>
<tr>
<td>4EBE</td>
<td>2005/03/15 11:07</td>
<td>2005/03/19 21:56</td>
<td>5</td>
<td>61.910%</td>
</tr>
<tr>
<td>4EBF</td>
<td>2005/02/28 10:09</td>
<td>2005/03/19 23:02</td>
<td>20</td>
<td>66.717%</td>
</tr>
</tbody>
</table>
Deployment #1

- Results (cont.)
  - Missing data largely correlated with sleep periods
  - Despite the low yield, promising data for activity model
    - Using elastic burst detection [Zhu and Shasha 2003]

**Lesson learned:**
Design should focus on improving yield
- Define degraded modes
- Improve on-mote storage
Experiment #1

Figure courtesy of Cecile Cornou
Deployment #2 (Feb 2007)

- Goal:
  - 3 weeks
  - 12 sows
    - Batch1: 6 sows
    - Batch2: 6 sows
  - Collect time series
    - 3D accelerometers
    - 4 Hz sampling
    - 90% yield
Deployment #2

- **Approach:**
  - Sensinode micro running TinyOS
    - 802.154 radio
    - 512 KiB flash
    - Accelerometer board
    - 2 NiMH cells (1.25V, 2400mA)
    - Compression
    - Control loops over sample, compress, store, send
  - Access points with external antennas

![Diagram with labels:]
- Data center: 1 TB
- Data and video repository
- Gateway
- Network monitoring
- Scientist analyses data and video
- Network monitoring
- Gateways
- Data and video repository
- 512 KiB/sec
- Data center 1 TB
- Scientist analyses data and video

Legend:
- R A = Resting Area (Straw bedded)
- A A = Activity Area
- F S = Feeding Station
- B P = Boar Pen
Deployment #2

- **Batch 1**

<table>
<thead>
<tr>
<th>mote</th>
<th>samples</th>
<th>received</th>
<th>%</th>
<th>clean</th>
<th>%</th>
<th>errors</th>
<th>promille</th>
</tr>
</thead>
<tbody>
<tr>
<td>33cb</td>
<td>6659775</td>
<td>6155871</td>
<td>92.4</td>
<td>5725802</td>
<td>86.0</td>
<td>10224</td>
<td>1.5</td>
</tr>
<tr>
<td>4228</td>
<td>6664812</td>
<td>6202083</td>
<td>93.1</td>
<td>5853400</td>
<td>87.8</td>
<td>13372</td>
<td>2.0</td>
</tr>
<tr>
<td>5d01</td>
<td>6659416</td>
<td>6316440</td>
<td>94.8</td>
<td>6295749</td>
<td>94.5</td>
<td>8842</td>
<td>1.3</td>
</tr>
<tr>
<td>5f21</td>
<td>6660201</td>
<td>6310652</td>
<td>94.8</td>
<td>6289002</td>
<td>94.4</td>
<td>9338</td>
<td>1.4</td>
</tr>
<tr>
<td>6902</td>
<td>6664218</td>
<td>6237412</td>
<td>93.6</td>
<td>5908038</td>
<td>88.7</td>
<td>13325</td>
<td>2.0</td>
</tr>
<tr>
<td>7a40</td>
<td>6667783</td>
<td>5947416</td>
<td>89.2</td>
<td>5358573</td>
<td>80.4</td>
<td>4963</td>
<td>0.7</td>
</tr>
</tbody>
</table>

- **Batch 2**

<table>
<thead>
<tr>
<th>mote</th>
<th>samples</th>
<th>received</th>
<th>%</th>
<th>clean</th>
<th>%</th>
<th>errors</th>
<th>promille</th>
</tr>
</thead>
<tbody>
<tr>
<td>472b</td>
<td>6525533</td>
<td>6236216</td>
<td>95.6</td>
<td>6196884</td>
<td>95.0</td>
<td>6373</td>
<td>1.0</td>
</tr>
<tr>
<td>4830</td>
<td>6693554</td>
<td>6336382</td>
<td>94.7</td>
<td>6310683</td>
<td>94.3</td>
<td>2438</td>
<td>0.4</td>
</tr>
<tr>
<td>4e78</td>
<td>6692651</td>
<td>6352667</td>
<td>94.9</td>
<td>6352204</td>
<td>94.9</td>
<td>6903</td>
<td>1.0</td>
</tr>
<tr>
<td>72ca</td>
<td>6694049</td>
<td>6434630</td>
<td>96.1</td>
<td>6394319</td>
<td>95.5</td>
<td>12599</td>
<td>1.9</td>
</tr>
<tr>
<td>73c9</td>
<td>5910142</td>
<td>5398960</td>
<td>91.4</td>
<td>5019267</td>
<td>84.9</td>
<td>11704</td>
<td>2.0</td>
</tr>
<tr>
<td>7c60</td>
<td>6694338</td>
<td>6230244</td>
<td>93.1</td>
<td>5722659</td>
<td>85.5</td>
<td>2279</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Lessons Learned

- Degraded modes
  - Kick in case of failure / pressure
  - Lossy mode based on how collected data is to be used

- Key mote characteristics
  - Amount of storage
  - Time during which the radio must be on
  - Power consumption in sleep mode
MANA (2008-)

- Lake Monitoring at Zackenberg (NE Greenland)
  - Manual sampling
  - Measuring biology parameters (Chlorophyll, dissolved O2, temperature, ...)
  - A few data points a year, over 10 years

MANA is a collaboration between DIKU, KU FBL, RU, Arch Rock Corp., Dan-Systems Aps.
Instrumentation

- **Buoy:**
  - Expensive sensor (WQM from Wetlabs)
  - Commercial mote-based solution
    - 6LowPan motes from Arch Rock with serial port
    - Low power listening
  - Mechanical design from Dan-System
  - 200 Ah Lithium battery pack

- **Base station**
  - Gateway Based on Microserver (PCS + SBC)
  - 6x rechargeable batteries (600 Ah) + solar panel
  - Autonomous Data Acquisition
Autonomous Data Acquisition

Rationale:

(a) Improve contingency planning (the system should not wait for a technician to react to errors or interesting events)

(b) Meet ecologists requirements (getting rid of impedance mismatch between ideal expectations and best effort implementation)

Approach: Three-layer autonomous system architecture
Lessons Learned

Extreme Instrumentation

Do not design for high availability. Failures will happen. Design for limited damage and easy repair.
Lessons Learned

- Benchmarking Flash devices is hard [Bouganim et al. 2009]
- Non-uniform write performance
- Performance depends on device state

Random Writes – Samsung SSD
Out of the box

Random Writes – Samsung SSD
After filling the device
Outline

Trends and Challenges

Where are we now?

Hogthrob: Characterizing Mote Performance

Mana: Autonomous Data Acquisition

Where do we go from here?
Perspectives

- Beyond Science, Physical Systems with new capabilities
  - Ecosystem monitoring, energy monitoring, regulation of car traffic, ...
- There are already many deployed, interconnected sensors (surveillance cameras, building automation, ...)
- What is new is ...
  - System's perspective on Sensornet-Based Instrumentation.
  - Generation of radically new data management architectures. The challenge is to efficiently support the sensor data lifecycle.