Thoughts on measuring and managing energy in HPC

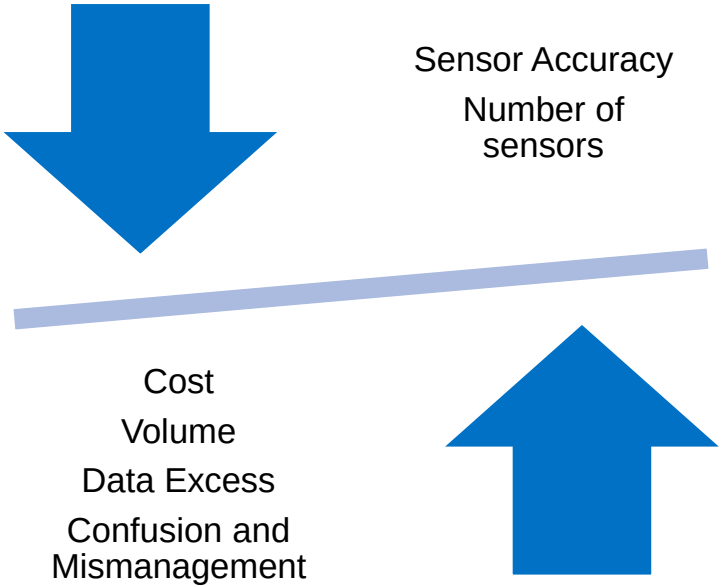
Michael K Patterson
Technical Computing Group
TCG System Architecture and Pathfinding

With help from Ram Nagappan, Jay Vincent, Nishi Ahuja

Energy/Power Management capabilities from Component to System

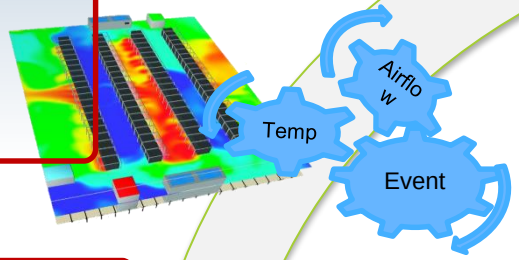
- Component
  - RAPL (Running Average Power Limit) to Monitor and Limit – CPU and Memory Power
    - counters are updated ~every millisecond
    - accuracy of the data around 10% at low power levels, better as we approach TDP
- Node
  - Intel Node Manager to Monitor & Limit Node Power
    - Sampled every 100ms
    - It is based on PMBUS power supply specification
      - Around +/- 5%
- System/Cabinet
  - Intel Data Center Manager (DCM) to monitor and limit power at System/Cabinet

An interesting balance

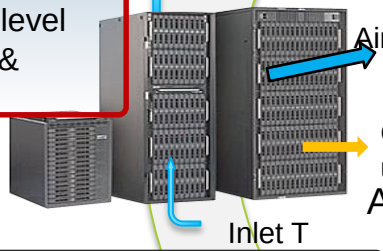


Power and Thermal Aware Solution Approach

Evaluation
Hot / Cold spot
Workload assessment
Bypass
Recirculation



Monitor
Rack and Row level data collection & aggregation



Telemetry Data

Platform-based sensors at each server vs. 3rd Party sensors in isolated locations

Respond

- **Cooling**
  - Adjust temp
  - Adjust airflow
  - Uniform temp
  - controls
- **Compute**
  - Headroom
  - Compute energy
  - Workload migration
- **Mitigation**
  - Alerts
  - Power policy

Benefits :

1. OpEx savings
2. Cooling CapEx savings
3. External sensor instrumentation savings
4. Extend capital lifespan

* Innovation opportunities

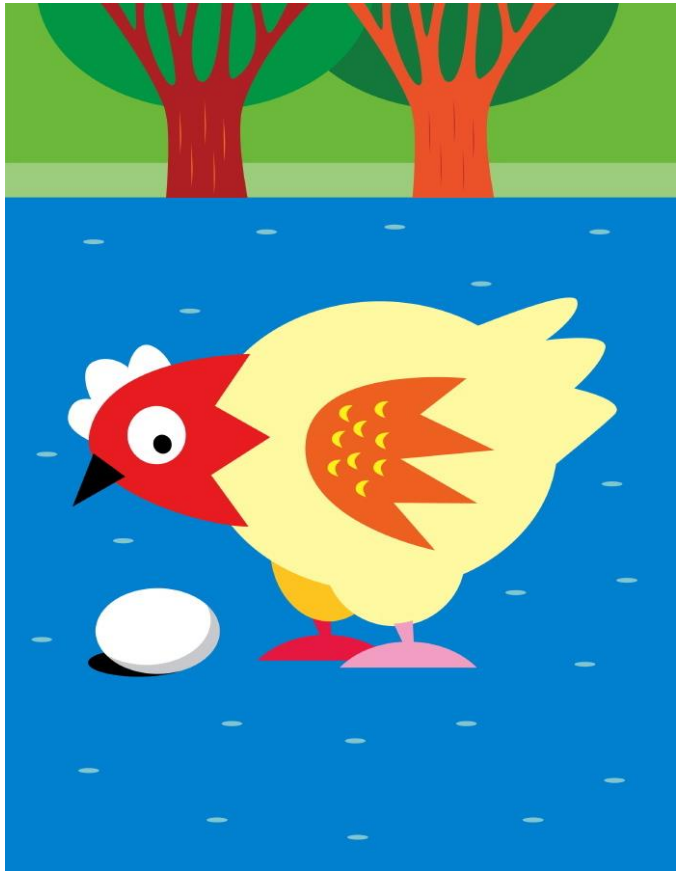
A conundrum….

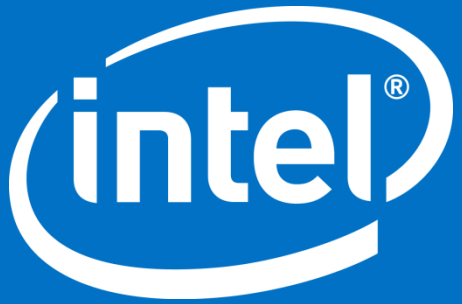
Current capabilities meets what's needed to get system level monitoring, job level reporting on energy use.

Next step? Application tuning for energy efficiency. Which is the impediment?

a) We can’t because we don’t have the sensors (number, speed, or accuracy)

b) We won’t because we get core-hour allocations, instead of kW-core-hours




intel®