From Sensors to Supercomputers
Big Data Begins with Little Data

ARM

Eric Hennenhoefer
VP Research

Linaro Connect 2016

Introducing ARM Research

- About ARM Research
  - 3 – 7 years ahead of product teams
  - From advanced development to blue sky
  - Locations in Austin, Cambridge UK, San Jose, and Shanghai

- Objectives
  - Build a pipeline to create and bring future technology into ARM products
  - Create and maintain the technology roadmap
  - Enable academia and research partnerships

Research Focus Areas

Memory & Interconnect

JEDEC

Tracking and Driving Memory Roadmaps

- Leading future task group

Hybrid Memory Cube CONSORTIUM

Going Beyond Evolutionary DRAM

- 3D stacked memories
- Intent-based interfaces

NVM in the System

- Drive technology
- Ensure open standards

Compute Near Memory

- Reduce data movement

Architecture

Embedded Efficiency

- TrustZone-M
- Improve code density and performance

Next Gen Arch

- Super secret stuff
- Use transistors more efficiently
- Accelerate key use cases

Security

- HW is the root of trust
- Make is easier to write secure SW

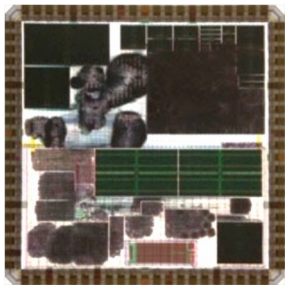
New Apps

- Novel use cases

ARM

Research Focus Areas

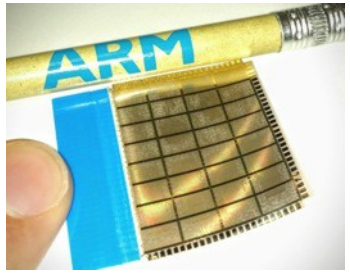
Applied Silicon



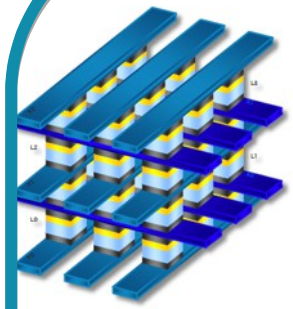
IoT Sensor Nodes
- Sub-threshold for 0.1x energy
- Energy optimized mixed-signal
- Extreme power gating

Integrating everything
- Voltage regulators
- Energy harvesters
- Sensor interfaces

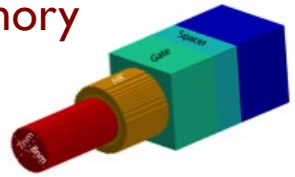
Printed Electronics
- 1cent disposable MCUs
- Mapping the ecosystem

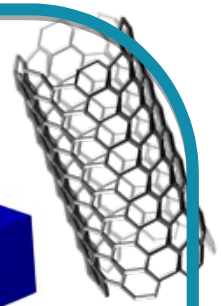


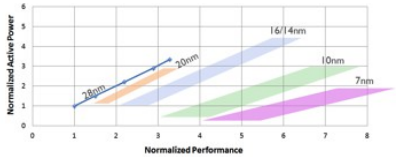
Future Si Tech



Disruptive technology
- Next Big Thing Memory
- What's after MOS?
- 3DIC technology

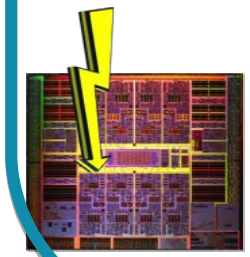






Predictive Technology Modeling
- Technology scaling entitlement
- Design-Technology Co-Optimization
- Next node device, patterning, ..

ARM layout on EUV. ASML ISSCC 2013.



Dependable Computing
- Detection, Correction, Security
- Robust power delivery

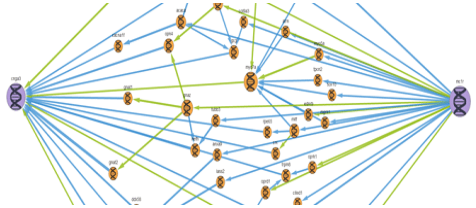
ARM

Research Focus Areas

Large Scale Systems

High Performance Computing

- Enable the first ARM supercomputer



Data Intensive

- Improving system efficiency for analytics workloads

sideARMs

- Compute near memory, network, and storage & standardize systems software interfaces

Design Integrity

Formal Methods

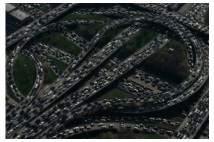
- Formal Coherency Verification on Cortex®-A

CPU µArch Models

- Verifying implementations against executable spec

Rain

- Deriving RTL checkers from Architecture specification



Deadlock Dependency Models

- Design-time deadlock freedom for arbitrary interconnect topologies

ARM

Research Focus Areas

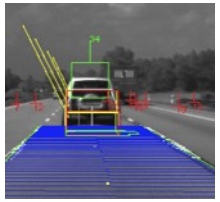
Emerging Applications

Machine Learning

- Speech & image recog
- Neural networks

Graphics Systems

- Full-system modeling
- System cache arch



Computer Vision

- Emphasis on automotive
- Depth perception, object and motion tracking

Mobile Systems

- Advanced workloads
- HW + SW system design
- Future devices

Special Projects

ARM motor

- Novel motor control

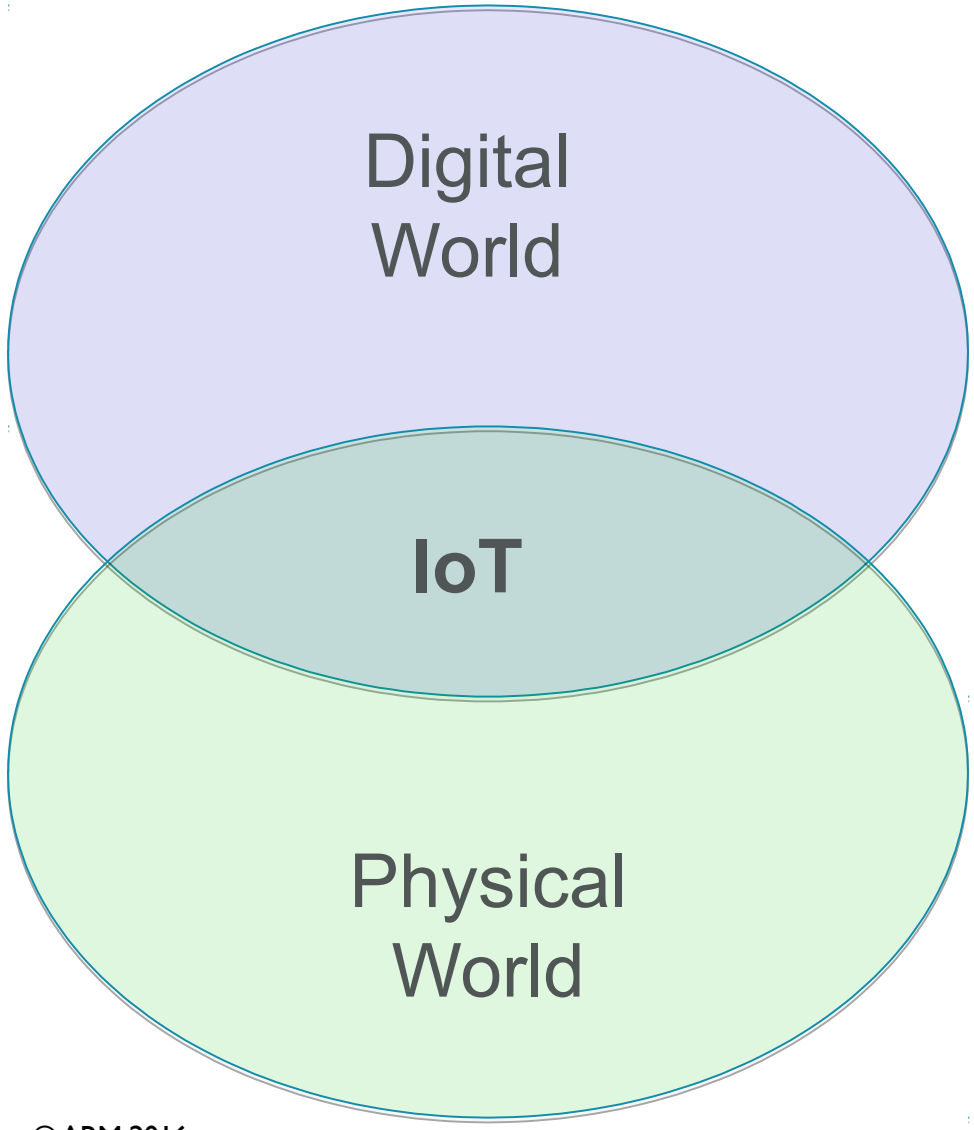
Technology Roadmapping

Technical Due Diligence

Low Power Radio

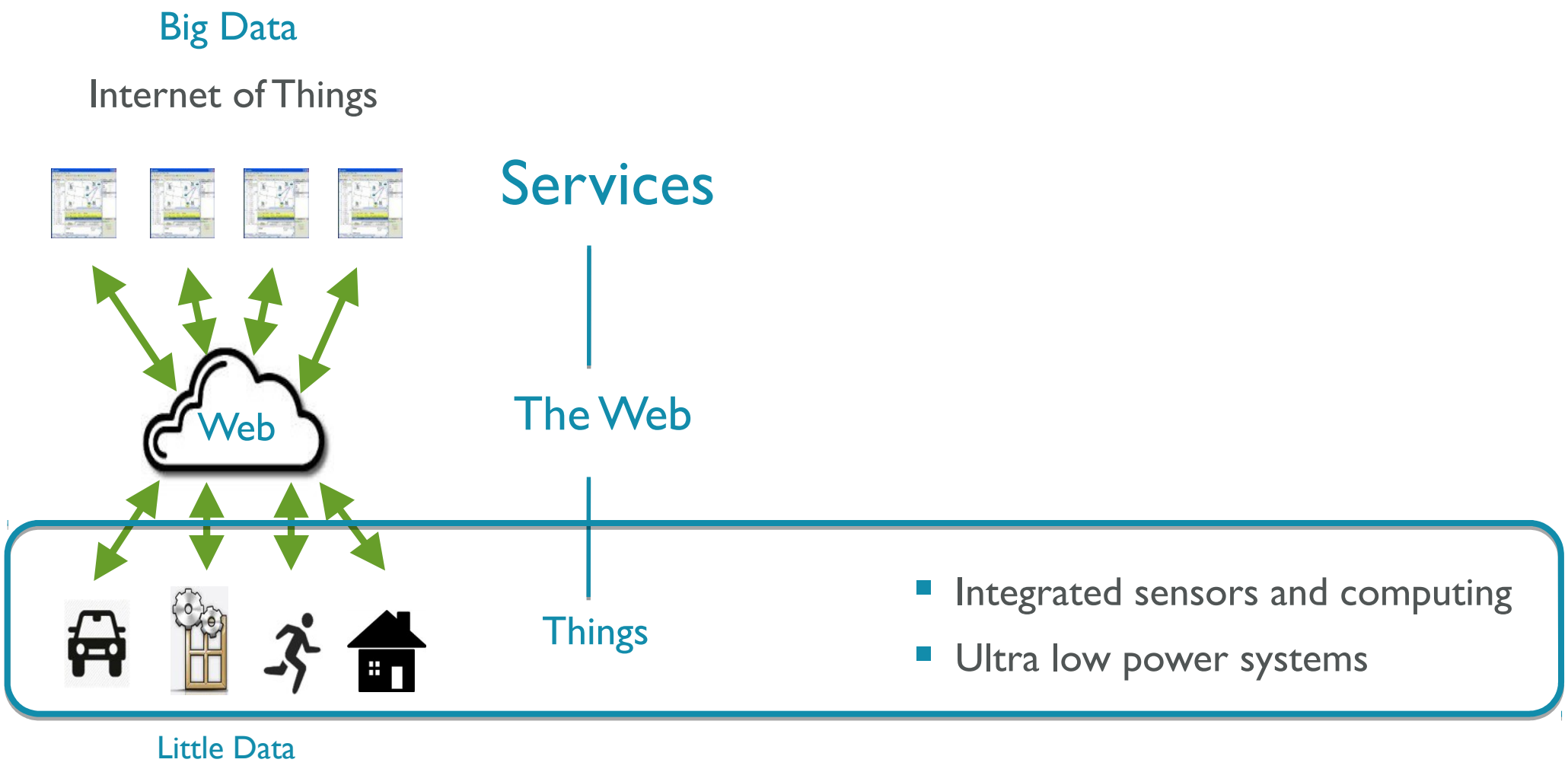
ARM

What Problem is IoT Solving?



ARM

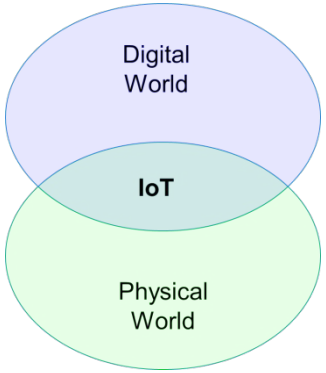
The key is in the connections: Hardware and Software



- Integrated sensors and computing
- Ultra low power systems

ARM

Sensors are the heart of the IoT



- Technically, the eyes, ears, nose, mouth and hands
- Sensors + compute + connectivity = IoT

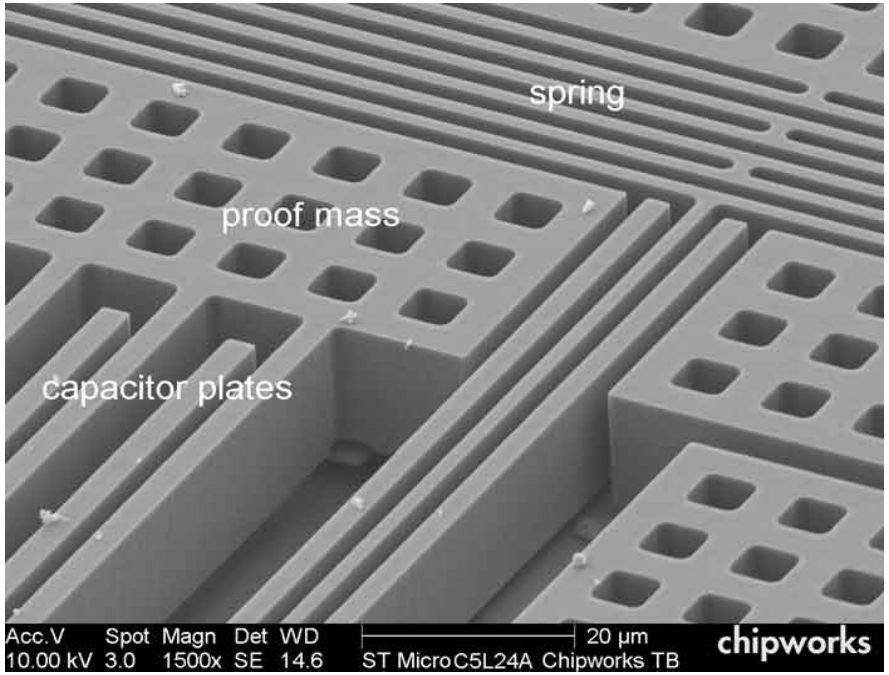
1T SENSORS IN 10 YEARS

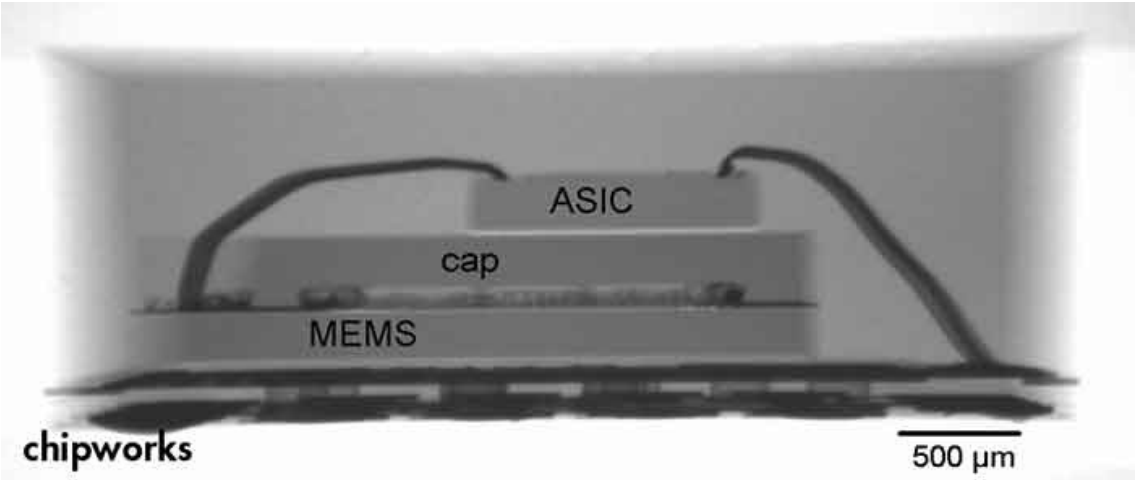
Year	Unit Price	Units Sold	Industry Revenues	Developed Population	MEMS Rev/ Person	MEMS Unit/ Person
2005	30.000	46,666,667	5,000,000,000	4,000,000,000	1.25	0.01
2010	15.000	466,666,667	7,000,000,000	4,000,000,000	1.75	0.12
2015	1.800	8,333,333,333	15,000,000,000	4,000,000,000	3.75	2.08
2020	0.216	138,888,888,889	30,000,000,000	4,000,000,000	7.50	34.72
2025	0.026	1,388,888,888,889	60,000,000,000	4,000,000,000	15.00	347.22

Chris Wasden at 2014 MEC, via semiwiki

ARM

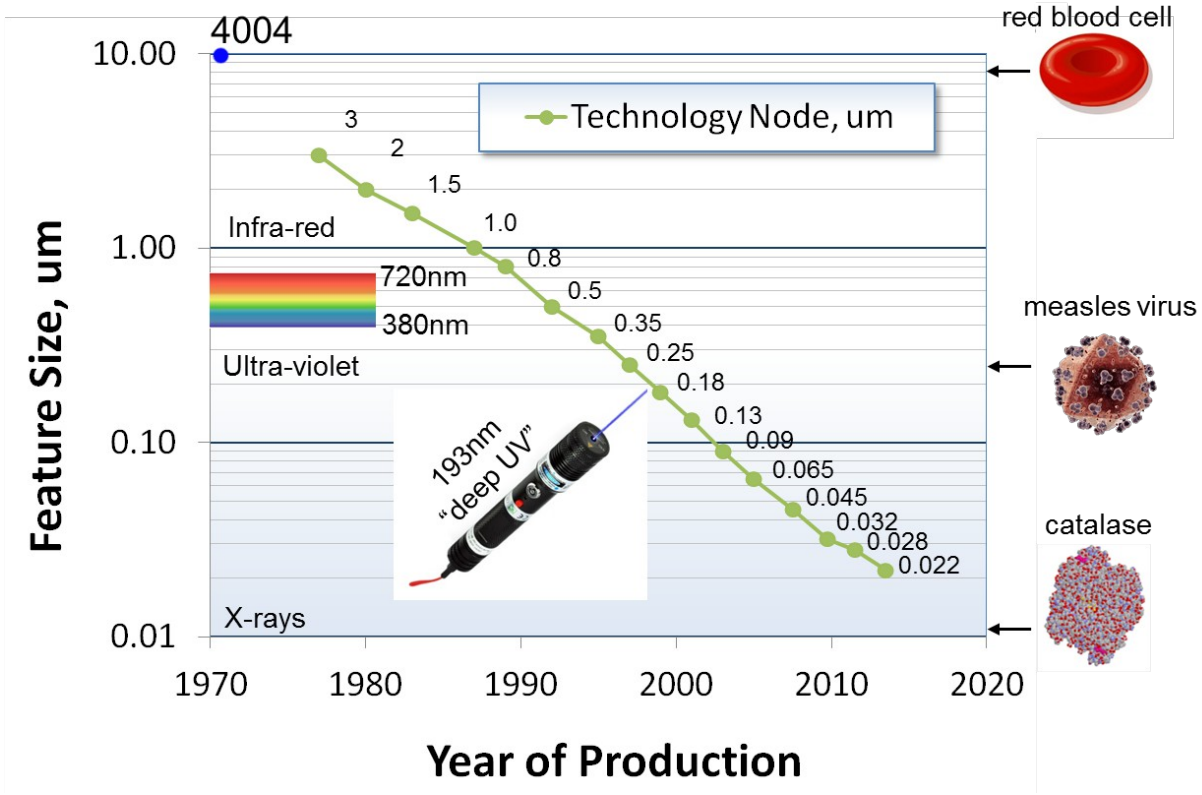
Sense of Touch: MEMS accelerometer, 3DIC

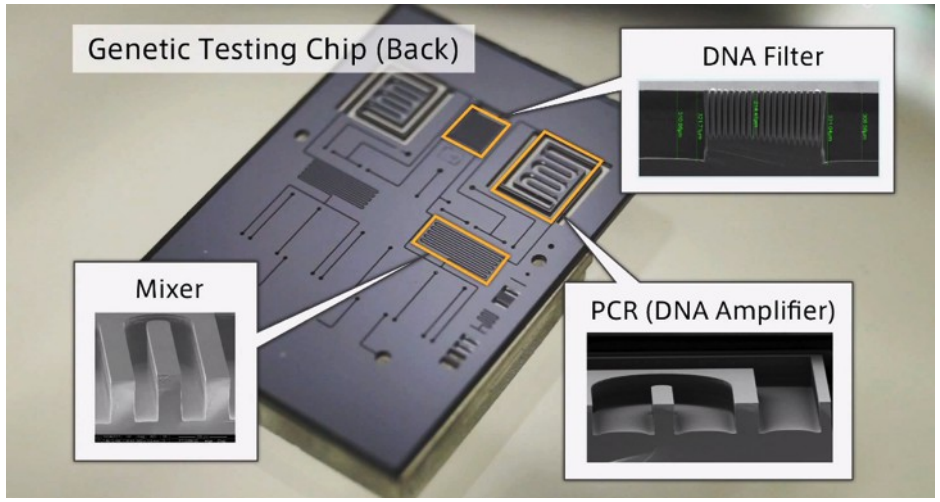




ARM

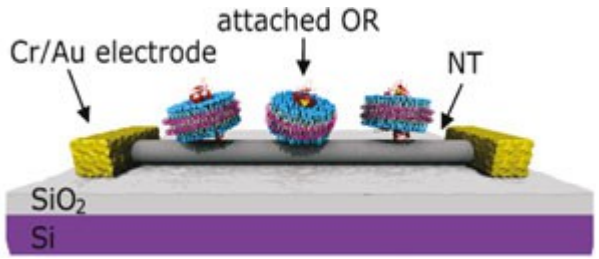
Feature shrinking: From cell-size to molecule-size



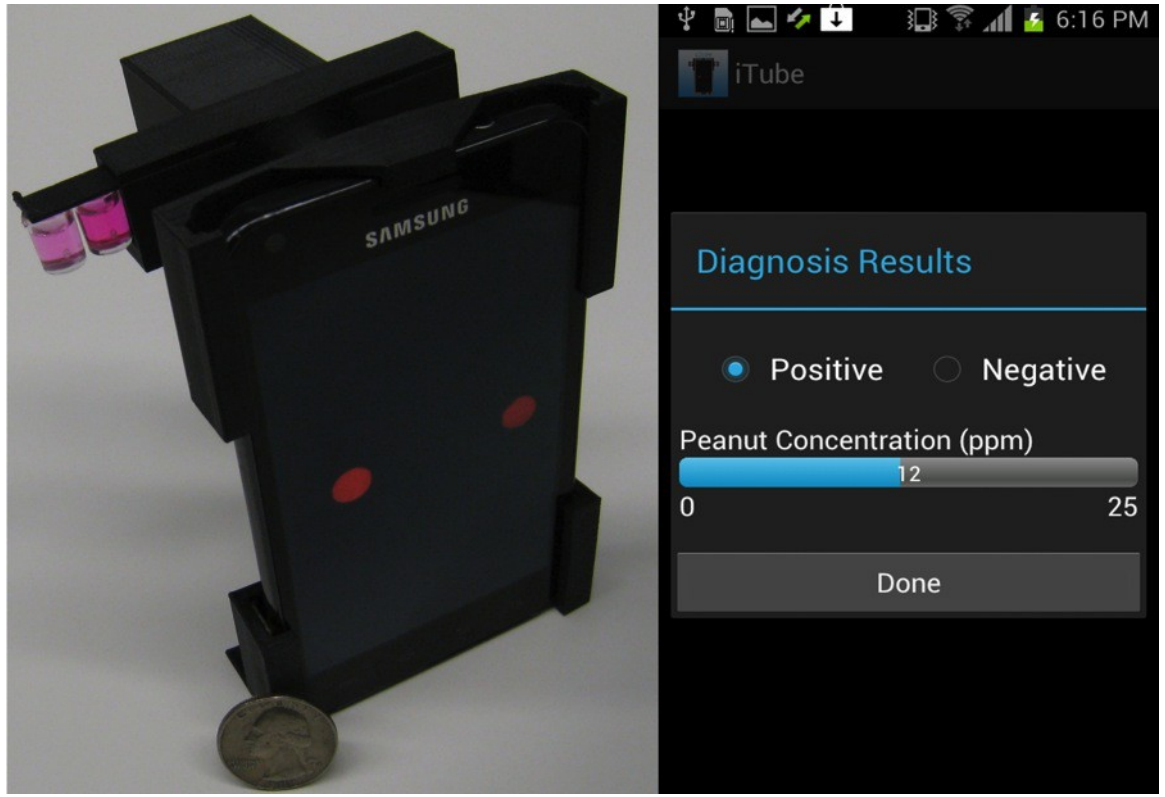


IMEC / Panasonic

Chip senses: Adding smell and taste



Adamant technologies, e.g.

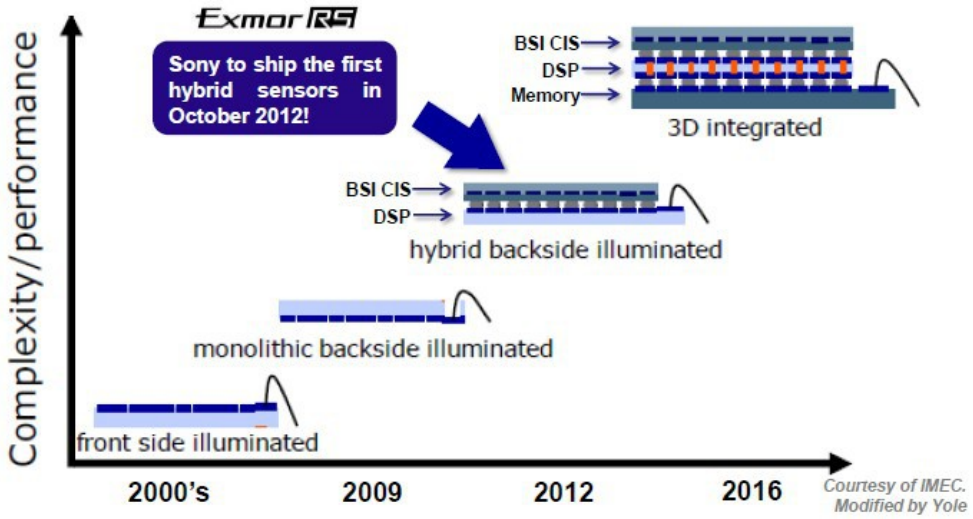


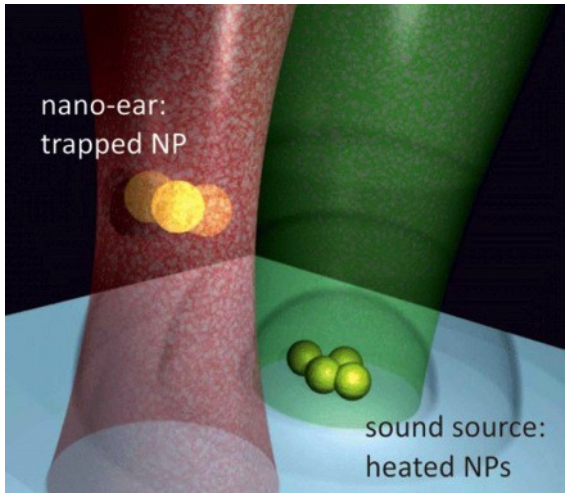
http://www.its.caltech.edu/~ahmet/publications.html

Food quality, air quality, infectious disease monitoring….

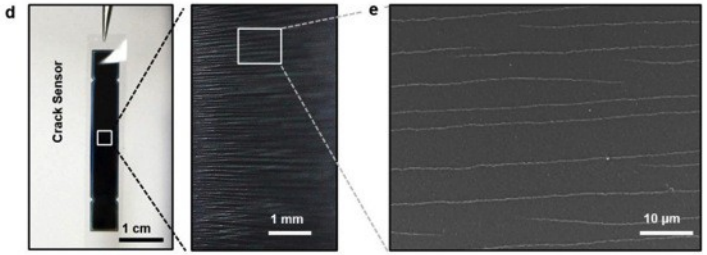
Adding hearing and sight

- Imaging – thank you cell phone industry
  - From the simple (is it daylight?)
  - To the complex (driving on the highway)
- Hearing and sight are senses that could be drastically augmented compared to our own





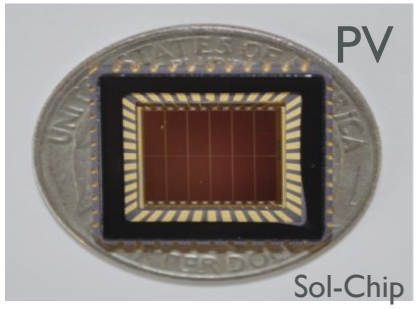
physicsworld



nature.com crack sensor

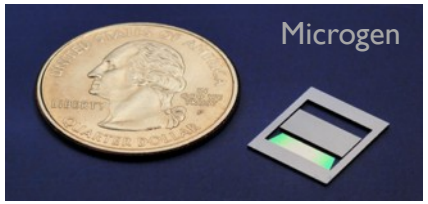
~~Sensors~~ Energy Harvesters will be the heart of the IoT

- Battery changing or re-charging is not "disappearing into the woodwork"
- Energy harvesters have low & very variable output power/voltage
  - And slow rate of improvement: ~1%/year for solar
- Nano-Watt standby allows bottom-end energy storage
  - Charging thin film battery or super-capacitor
- Minimizing peak power can reduce need for storage
  - Smaller & cheaper device
  - But not always available



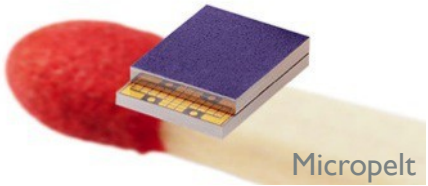
3mW in direct sunlight
20uW under office lights

Piezoelectric

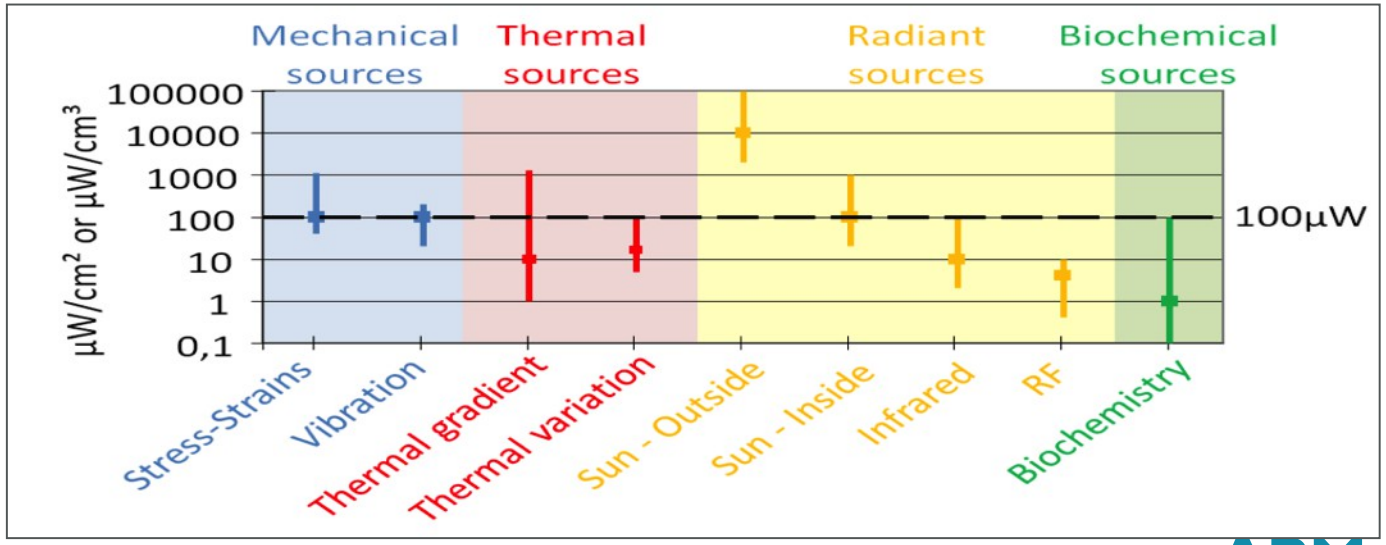


100uW @ 100Hz

Thermoelectric



300uW @ 5K

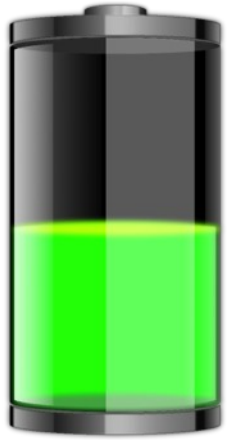


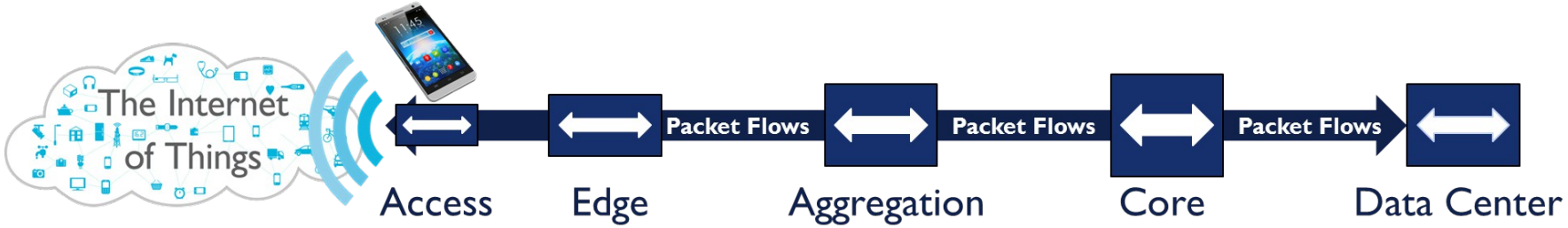
Theoretical energy density (Source: S. Boisseau, G. Despesse, and B.A. Seddik, "Electrostatic Conversion for Vibration Energy Harvesting," ArXiv e–prints, Oct. 2012.)

ARM

Energy Efficiency: Things you can do with 100pJ

- Run a Cortex®-M0 for 10 cycles
- Write one bit of flash
- Write ~300 bits of DRAM or SRAM
- Send ~5 bits across LPDDR4
- Transmit 2 bits of UWB data
- Transmit 0.02 bits over Bluetooth LE
- Drive an electric car 100fm (@1MJ/km) ~0.05% of the distance across Si atom





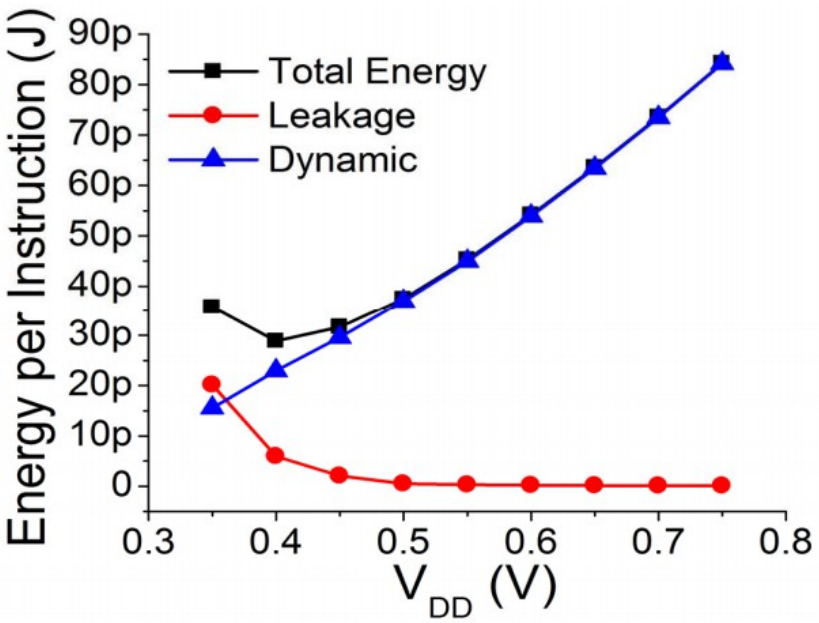
Energy costs to transmit, compute, and store data will define the shape of the IoT
VSLI Technology advancements will re-write the boundary conditions

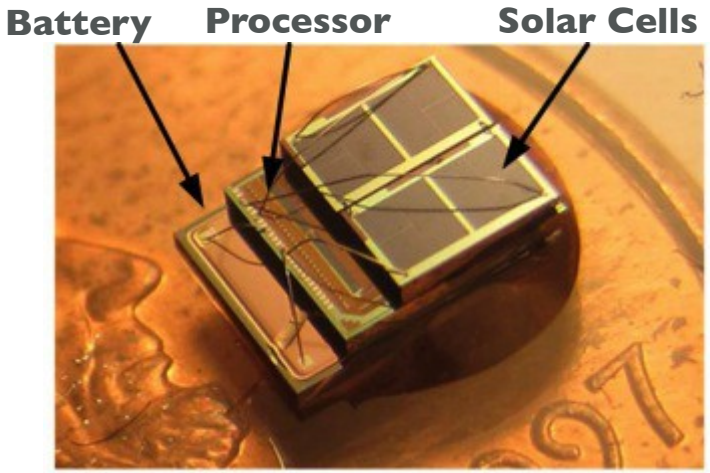
ARM

Or 1,000,000x / sec with 100 uW

180nm Mich Micro Mote: 30 pJ / cycle

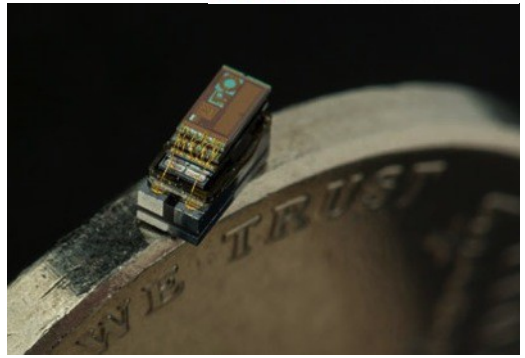
- Michigan Micro Mote
  - Cortex®-M3
  - 180nm, 8.75 mm^3
  - Vdd = 0.4V, Vt = 0.4V
  - 73 kHz/1 MHz operation





G. Chen et al., ISSCC, 2010.

Computer History Museum



http://www.eecs.umich.edu/eecs/about/articles/2015/Worlds-Smallest-Computer-Michigan-Micro-Mote.html

ARM

65nm M0+ : 11.7 pJ / cycle

8.1: An 80nW retention 11.7pJ/cycle active sub-threshold ARM Cortex®-M0+ sub-system in 65nm CMOS for WSN applications

James Myers, Anand Savanth, David Howard, Rohan Gaddh, Pranay Prabhat, David Flynn

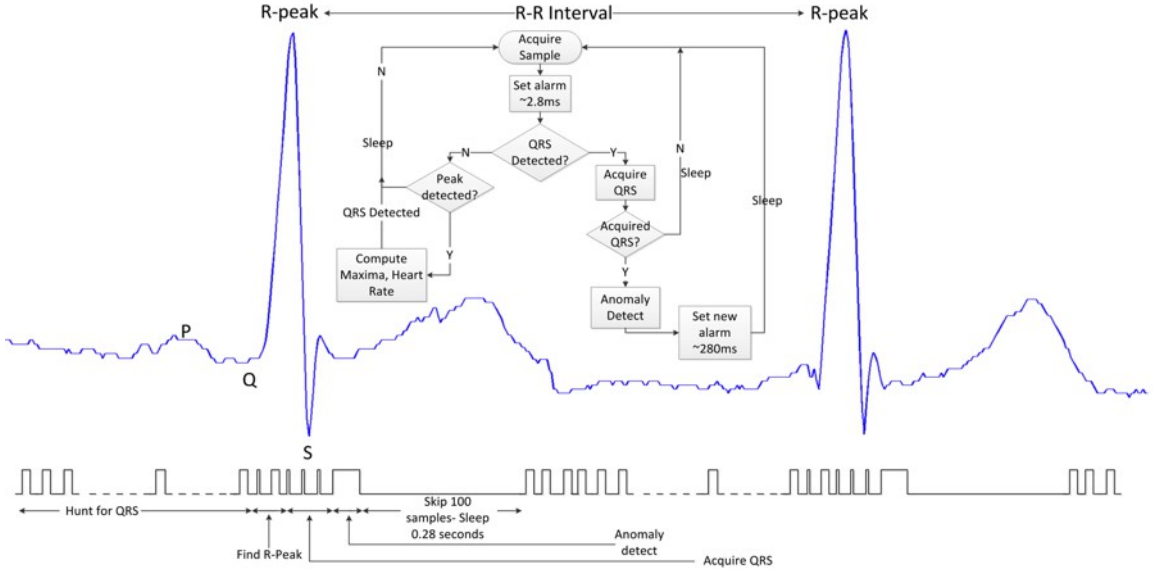
ARM®

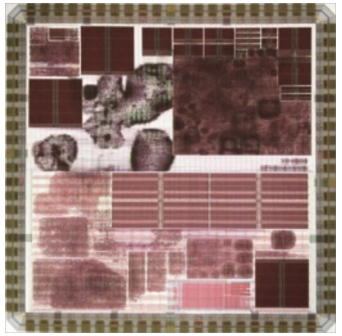
8.1: An 80nW retention 11.7pJ/cycle active sub-threshold ARM Cortex-M0+ sub-system in 65nm CMOS for WSN applications

1 of 25

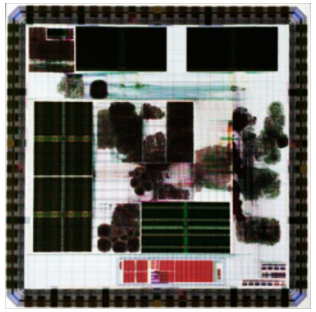
ISSCC 2015

Heart Monitor Workload





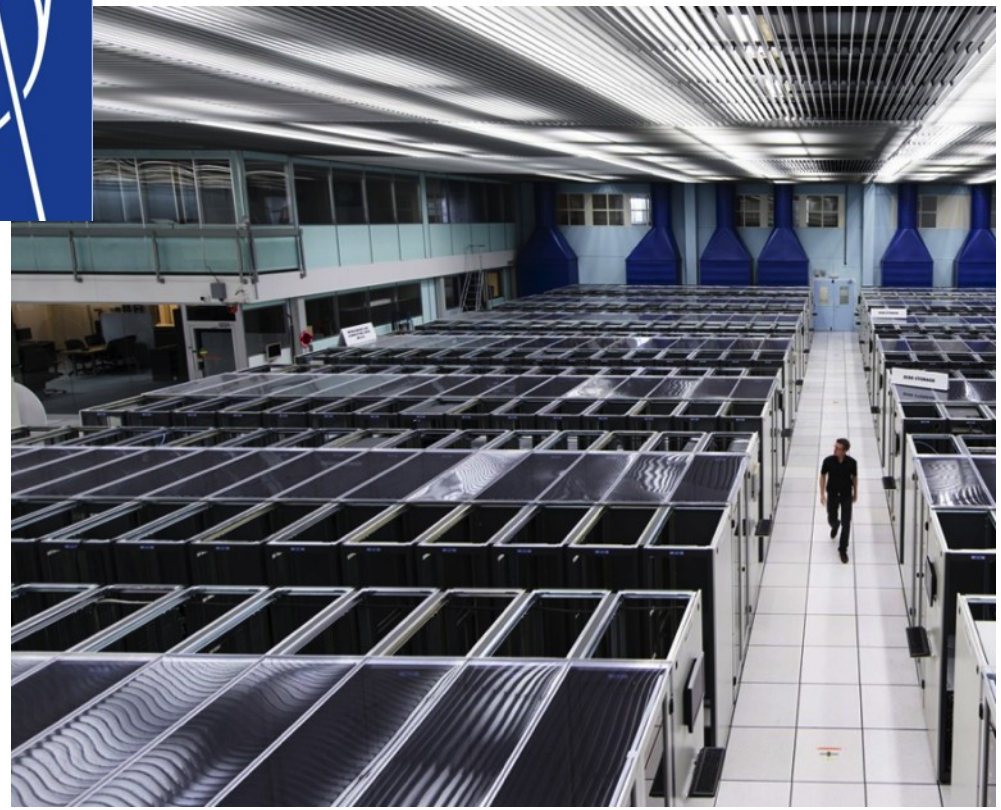
3.0μW



2.3μW

ARM

Big Science Starts With Little Data Too



Run2: 25GB/s

Transfer antennas to DSP: 200 TB/s

Streaming Data is the Next Challenge

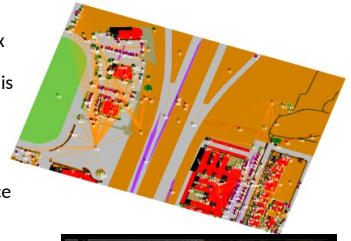
Geospatial Graph Analysis

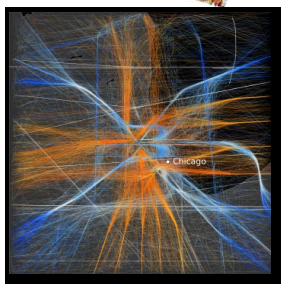
Reality
- Sensors produce enormous quantities of complex data
- Current analysis capabilities fail to fully exploit this data to produce actionable intelligence

Challenge: leverage the structure of geospatial data to identify patterns of life
- Automated data analysis capabilities that enhance human decision-making
- Scalable analysis over disparate temporal and geospatial scales
- Pattern analysis of complex trajectories

R&D is required at all levels of the software/ hardware stack to automate the capture, fusion, and analytics of geospatial data streaming from heterogeneous sensors.

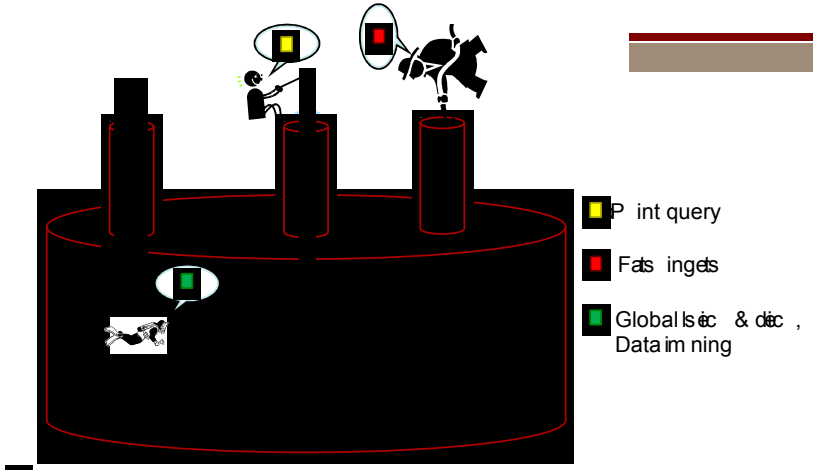
A convergence of HPC and graph-analytics is necessary to provide time-sensitive, actionable intelligence





2

Big Data Graph Streaming



1

Sandia National Laboratories

CCR Center for Computing Research

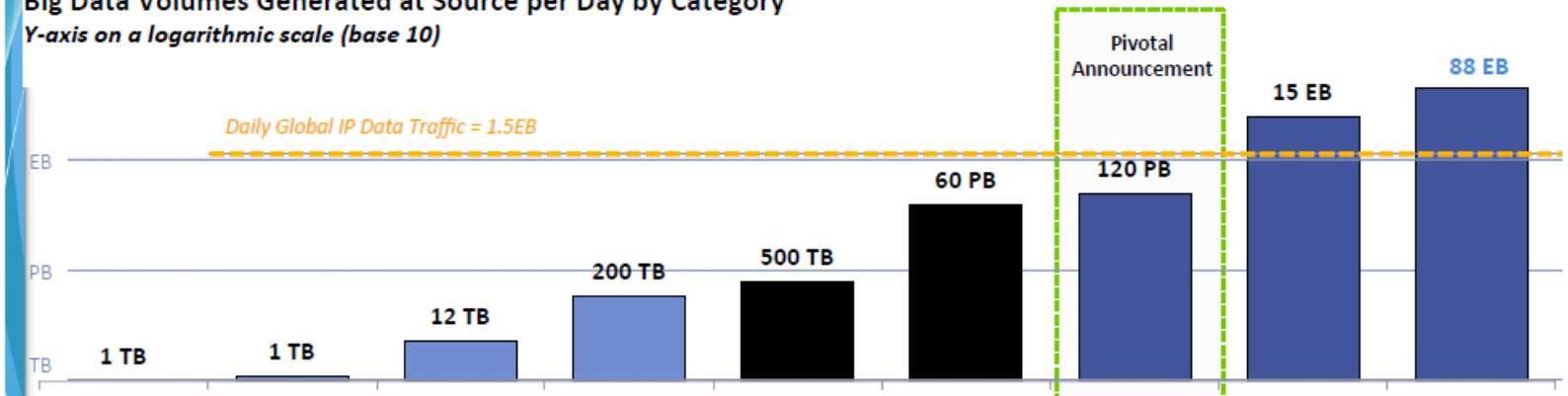
ARM

Even Basic Little Data can Produce a LOT of Big Data

IoT Big Data Generators Will Dwarf Platform Big Data Sources Like Video and Social!

Big Data Volumes Generated at Source per Day by Category

Y-axis on a logarithmic scale (base 10)



	NYSE	VISA	twitter	YouTube	facebook	Walmart	GE	BOEING	
	Financial		Social / UGC			Transaction	Sensor / Machine		
Drivers	• Total data generated from one day of trading	• 150M transactions processed per day • Assumed each transaction = 10KB of data	• 500 M users producing ~340 M tweets per day = 12 TB of data per day	• ~1 hours of new video uploaded every second • 1 min of video = 40 MB	• 950 M active users • ~2.5 B pieces of content generated per day	• ~1 M customer transactions logged per hour (2.5 PB / hour)	• 12,000 gas turbines WW • 20 sensors / turbine • 500GB / sensor/day	• 10 TB of data generated 30 mins of flight time • 4- engine jet = 640 TB of data from 1 trans-Atlantic flight • ~25,000 flights per day	• Up to 1 PB of data generated per second during experiments

Assumed each object on AWS S3 is 500KB in size

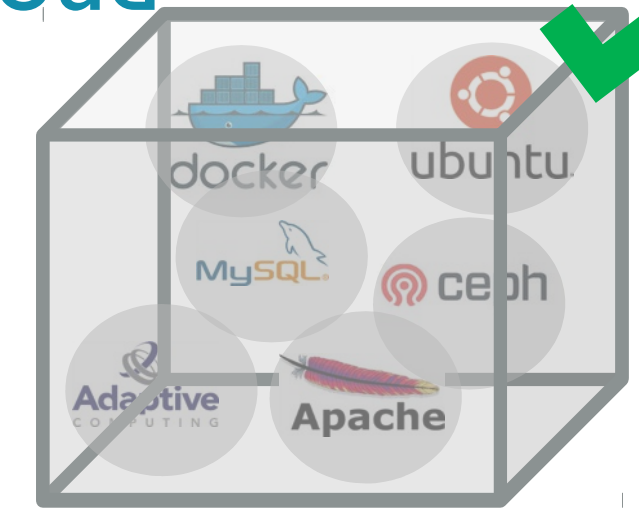
Flavio Bonomi, 2013

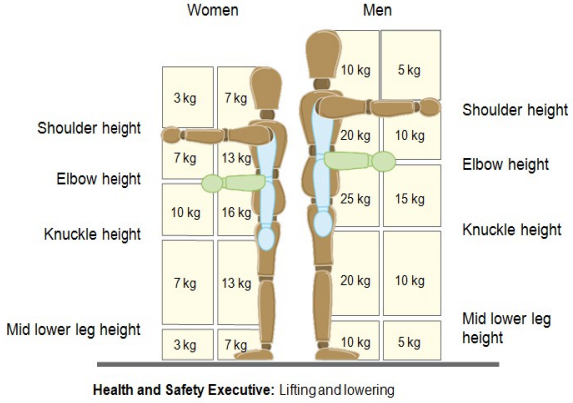
ARM

University of Cambridge Portable Cloud

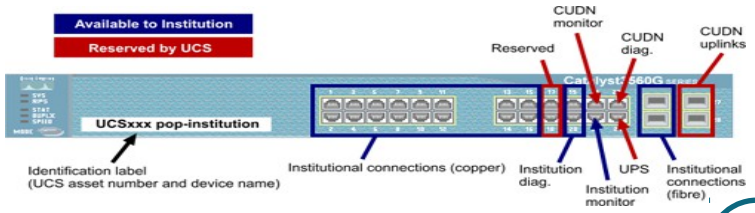
Power
2kVA

'Micro' data centres will dynamically adapt to support storage, web and computation.





Weight
50kg



Specification
(8x) 64 Servers
(8x) 256 Cores
(8x) 128TB Storage

UNIVERSITY OF CAMBRIDGE

ARM

Connected Teddy Bears: What Could Wrong?

- Hackers love IoT
  - If software hacks fail then
  - They will come via UART…
- The Basics – plan to be hacked
  - Harden the Device
    - Secure boot, Secure kernel, …
    - Perimeter security is insufficient
  - Intrusion detection
    - Deny foothold
    - Revert to known state
  - Secure over-the-air firmware updates
    - Plan to be hacked …

ARM

Connected Teddy Bears: What Could Wrong?

Engineering Scope [Security]

End to end security; from a light bulb to the cloud

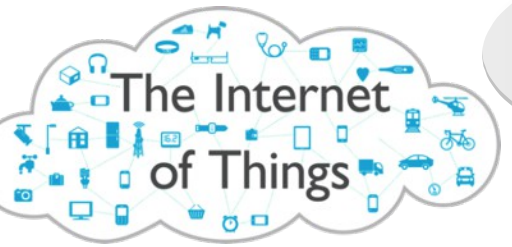
- Secure boot
  - Securely stored and authenticated digital signatures
- Access control
  - OS role based checking against authorised privileges
- Device authentication
  - Device needs to be authorised to access cloud services via the network
- Firewall
  - Device needs to be resilient against attacks
- Updates and patches
  - Over the air updates, functionality and security

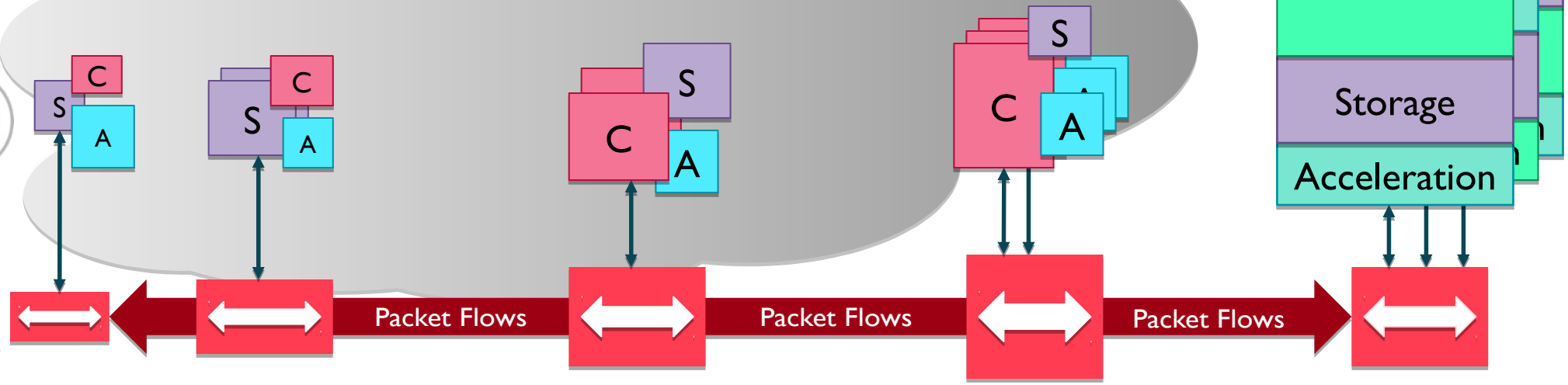
Linaro

ARM

The Journey From Little Data to Big Data

- Applications run where the data is, independent of the network node
- Heterogeneous Compute is distributed into the network
- Networks and Compute resources are both managed and configured using standard IT technologies





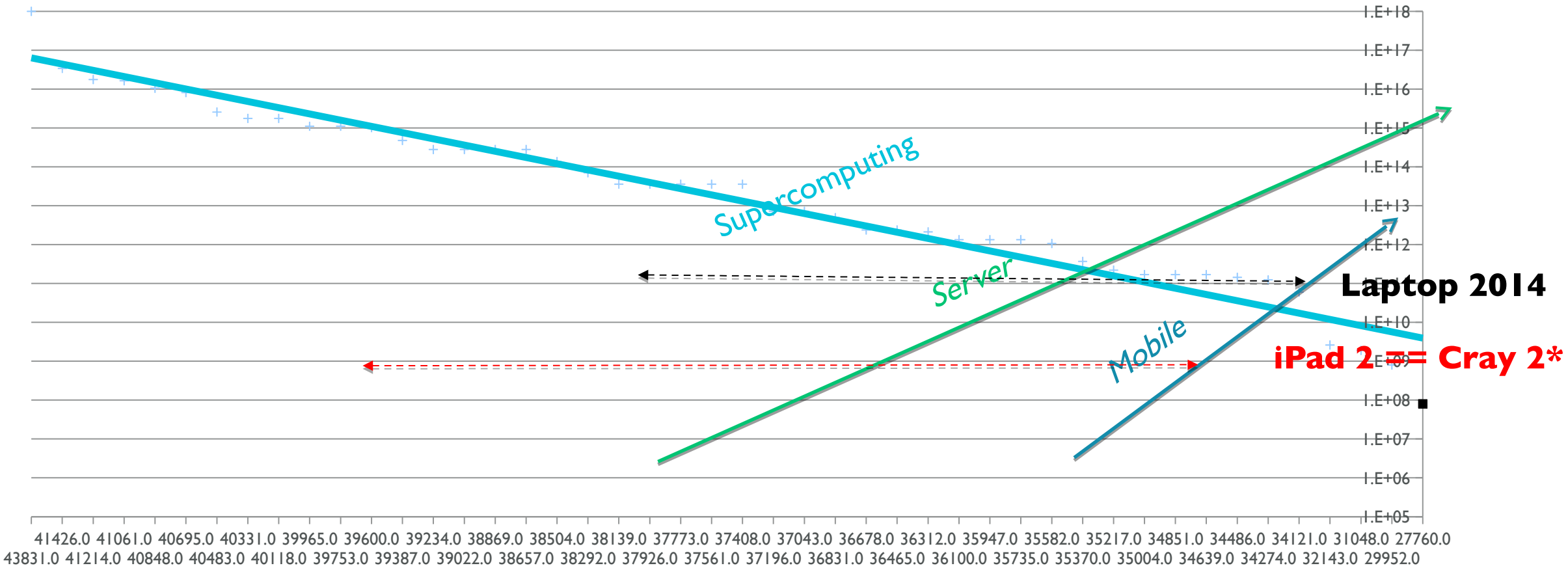
Scale-Down Power Consumption and Form Factor

Scale-Up from Little Data to Big Data

Decrease Latency

ARM

Why is ARM interested in Supercomputing?



* J. Dongarra & P. Luszczek HPEC 2012

ARM

High Performance Compute (HPC) – Why?

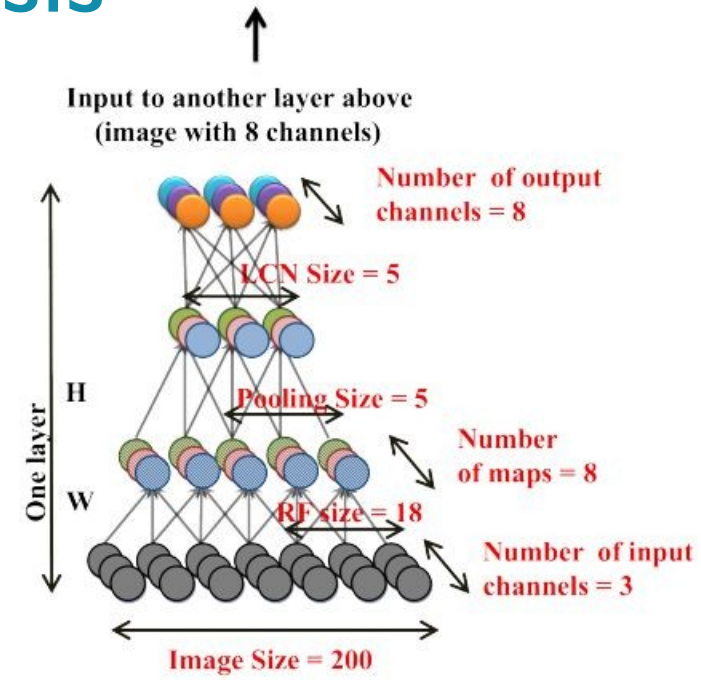
- Why ARM? - HPC community wants multivendor options
  - Strategic requirement
  - ARM ecosystem brings choice and a path to better optimized solutions

- Why Now? – Exascale is a compelling event
  - Massive parallelism is requiring changes to software, this opens the door for a new ISA
  - ARM HPC projects are active in multiple regions

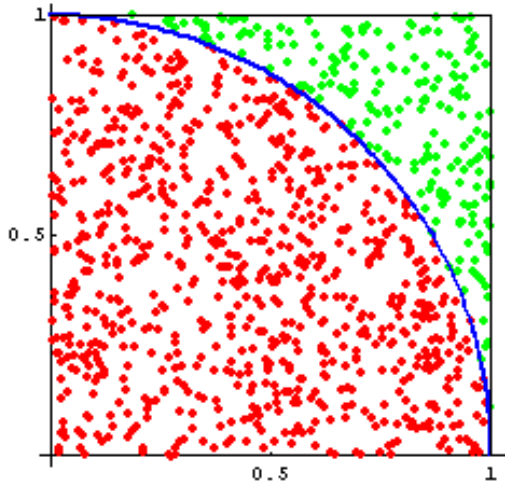
- Why Linaro?
  - HPC has a large open source component
  - Some customers require multiple tools chains: proprietary + open source

HPDA – High Performance Data Analysis

- 23% of HPC system usage is currently HPDA
  - Machine learning
  - Stochastic modeling / Monte Carlo – explore large problem spaces
  - MapReduce/Hadoop, graph analytics, knowledge discovery
  - Many fields benefit from real time results – finance



- Worldwide are migrating to commercial compute servers



Deep Learning – HPC is the Future

"This is why around 2008 my group at Stanford started advocating shifting deep learning to GPUs (this was really controversial at that time; but now everyone does it); and <u>I'm now advocating shifting to HPC (High Performance Computing/Supercomputing) tactics for scaling up deep learning</u>. Machine learning should embrace HPC. These methods will make researchers more efficient and help accelerate the progress of our whole field".

Andrew Ng - Quora Feb 3rd 2016

ARM

HPC Expectations: Platform Optimized Solutions

- Machine Learning on ARM example – 80% is about the Math(s)*
  - 1.0x ATLAS from repo is (single core)
  - 2.7x OpenBLAS from repo
  - 6.7x ATLAS self tuned (several hours setup)

- HPC expectations
  - Easy to access precompiled and optimized packages
  - Scientific packages: Compliers, MPI, math libs, profilers, schedules, pre-build python, …
  - Ability to make power trade-offs
  - Tuned for each silicon vendor and Linux distro

- Who will lead? OpenHPC? Linaro?

*Inference of AlexNet on Caffe with Batch-size 1, 8-core A57

ARM HPC Summary

- ARM HPC systems are coming, test beds are deployed
- Tool chains, apps, math libraries, are underway…
- Open source is a key component of HPC
- IoT is today and HPDA is a critical piece of the workloads of tomorrow
- From Sensors to Supercomputers: Big Data Begins with Little Data