18-640 Foundations of Computer Architecture

Lecture 23:
"Mobile & Cloud Computing: Mega-Trends and Future Directions" (A Personal Perspective)

John Paul Shen
Nokia Fellow

December 2, 2014

Electrical & Computer ENGINEERING
Carnegie Mellon University

AD 2013: Quite A Momentous Year…

ABSTRACT:

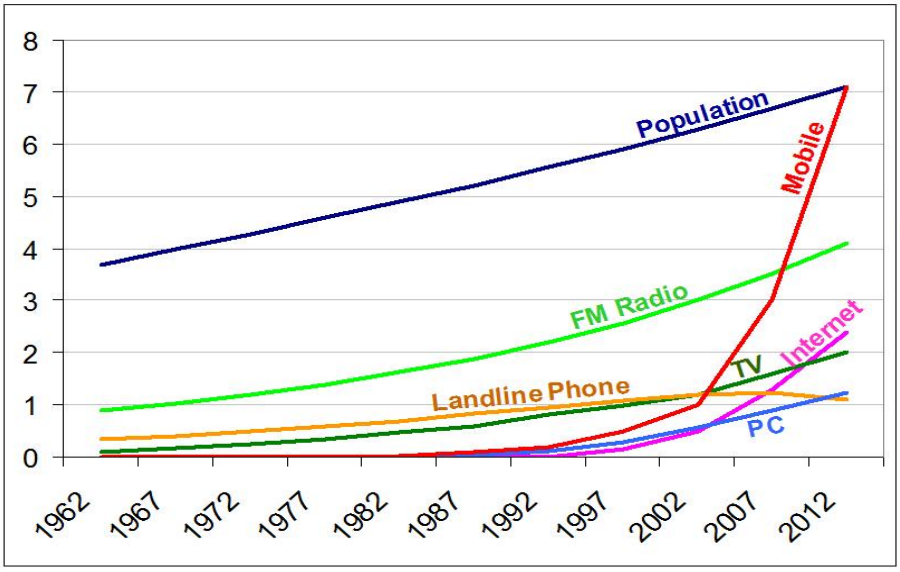
In the summer of 2013 the total number of mobile phone subscriptions world wide surpassed the total population of our planet.

> *"This is totally unprecedented in the human history of technology. No technology ever, has even come close. Not television sets, not PCs, not radios, not cars, not motorcycles, not even bicycles; not credit cards, not even bank accounts; not books in print, not newspaper circulations; not the reach of electricity or landline telephones or even running water; not wristwatches, not toothbrushes, not even pens and pencils… have been as widely used as mobile is today."* [Tomi Ahonen, 2013]

By the end of 2013 the total number of smartphones world wide exceeded the total number of PCs, of any kind, in the world. These smart mobile devices are mobile computers; they also serve as rich sensing platforms on the global scale.

This talk will highlight some of the research work done by Nokia Research - North America Lab on *Mobile Sensing and Social Services* and *Visual Computing and Mixed Reality*, and suggest some interesting directions for future research in the areas of *Human Behavior Modelling and Understanding* and *Global-Scale Mobile and Swarm Computing*.

Global Media End of 2012

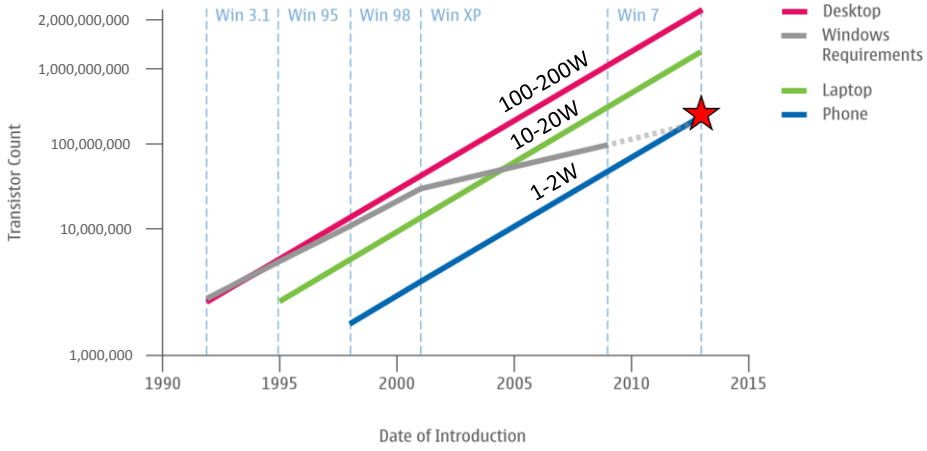


Source: *TomiAhonen Almanac 2012* and *TomiAhonen Mobile Forecast 2012-2015*

3

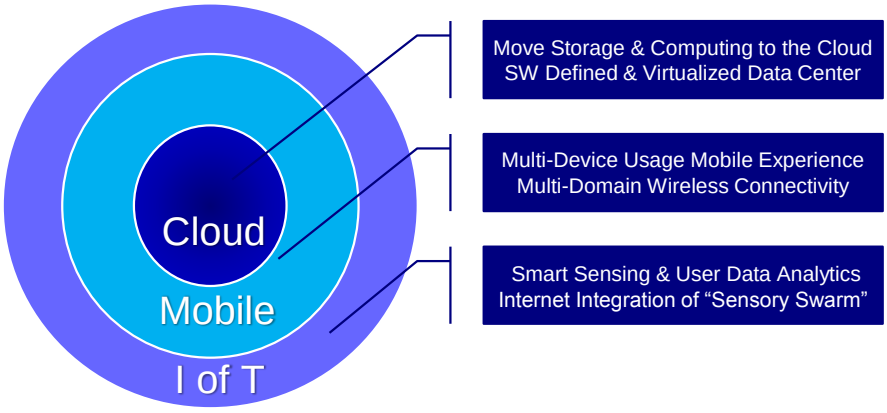
Moore's Law is Alive and Well

Cross Over Point in 2013



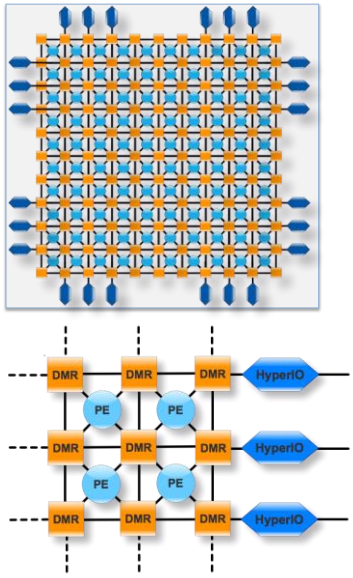
Current Mega-Trends

The computing cloud ecosystem is maturing and several trends are becoming evident and dominant



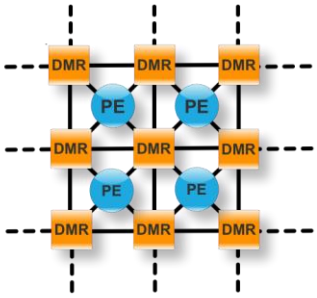
hx3100 Processor Based on the HyperX Architecture

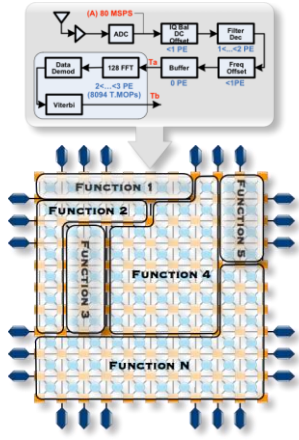
- 100 Processor Elements (PE)
  - 4 kB program memory per PE
  - 500 MHz variable clock
  - 8, 16, and n-by-16 integer capability
  - 32-bit floating-point capability
  - 50,000 MIPS, 50 GMACS 16-bit
  - 32-bit floating-point, 25 GFLOPS
- 121 Data Memory Routers (DMR)
  - 8 kB data memory per DMR
  - 8 DMA engines, high speed autonomous data routing
  - 968 kB total configurable on chip memory
  - Real time algorithm topology switching
- High Speed I/O Routers (HyperIO)
  - 24 Total HyperIO Channels, 16 bit wide
  - DDR2: 8 channels, 64 Gbps
  - LVDS: 12 channels, 96 Gbps
  - CMOS: 12 channels, 24 Gbps
  - 168 Gbps simultaneous, Access to 64 GB off-chip
  - I/O configurations are package-specific
- Low Power
  - 25 mW to 2.5 W typical core power
  - Automatic resource power management
  - Chip power-off by quadrant
  - 32 GFLOP/s/W (32 bit)
  - >150 GOP/s/W (16 bit)
  - >2.4 TOP/s/W (16 bit RISC equivalent)



Functions to Fabric

- A compute Fabric of 100 processors within a network of 121 data memory routers
- User configurable, adaptable in real time
- Supported by high bandwidth intercommunications for efficient data movement
- Each DMR contains 8 kB data memory + 8 DMA engines for neighborhood or cross chip data transfer





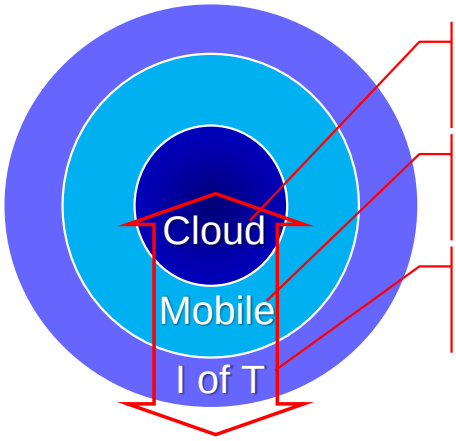
- Allows the appropriate topology to be created to support the natural parallelism of the algorithm / system
- without stalling resources in use
- and only clocking or powering what is needed
- … thus creating a highly efficient, low power system

Seamless Roaming Experience Across 5+ Screens

	PC	smartphone	tablet	living room	dashboard
	Mac computers	iPhone	iPad	Apple TV	CarPlay
Google	Chrome browser	Android	Android tablets	Google TV, Chromecast	Open Automotive Alliance
Microsoft	Windows, Office	Windows Phone	Windows 8	Xbox	Windows in the Car
amazon	Online store	Soon to come?	Kindle Fire	Fire TV	?

Potential Disruptions

The current cloud architecture can and will be disrupted as players begin to create new and better consumer experiences

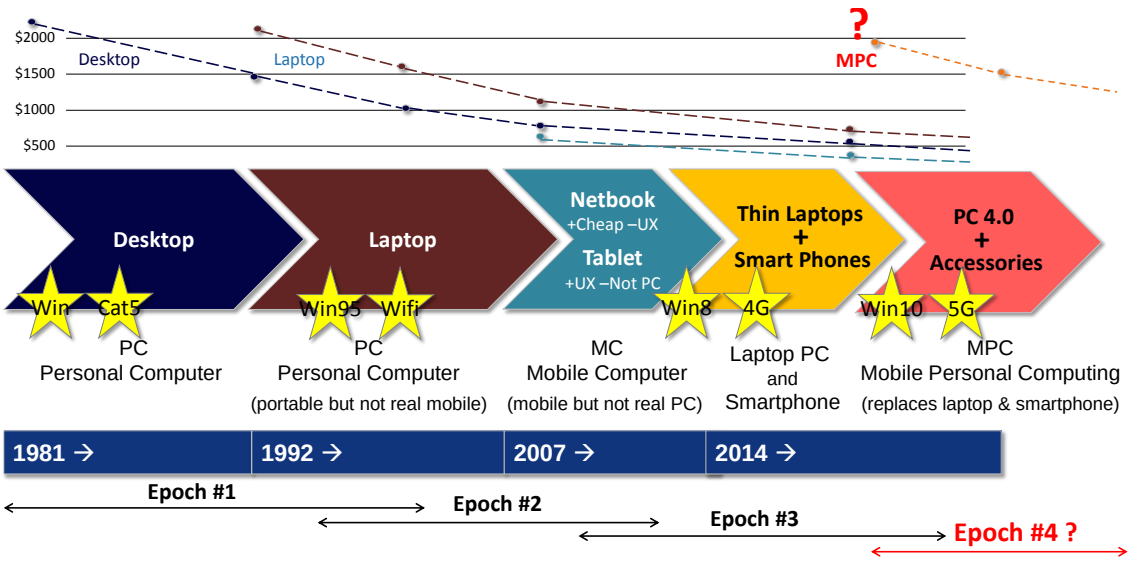


Shift Computing to the Cloud Edge:
Off Load Core Network Bandwidth Demand
Reduce Service Delivery Latency to Users

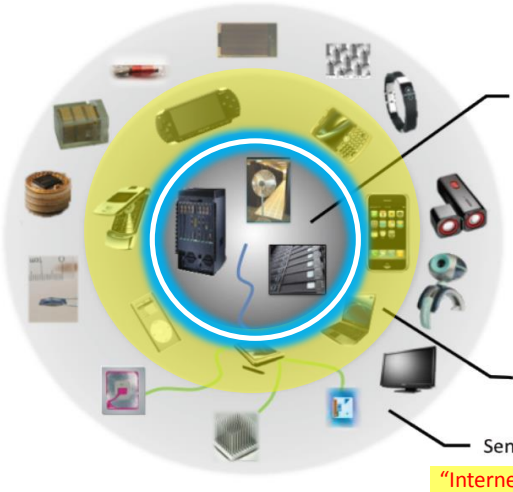
Truly Seamless Mobile Experience:
Seamless Cross-Device Cross-Domain UX
Unify both Broadband and Broadcast UX

Human Sensing for Common Good:
Deliver Real IOT Value to Mobile Users
Use both Eulerian and Lagrangian Sensing

New Epoch of "Mobile Personal Computing" (MPC) ?



Mobile and Cloud Computing Research Ideas



"The Edge of the Cloud"
Santa Clara, August 27, 2010
A Workshop Offered by the MuSyC and GSRC FCRP Research Centers

Electrical & Computer ENGINEERING
Carnegie Mellon University

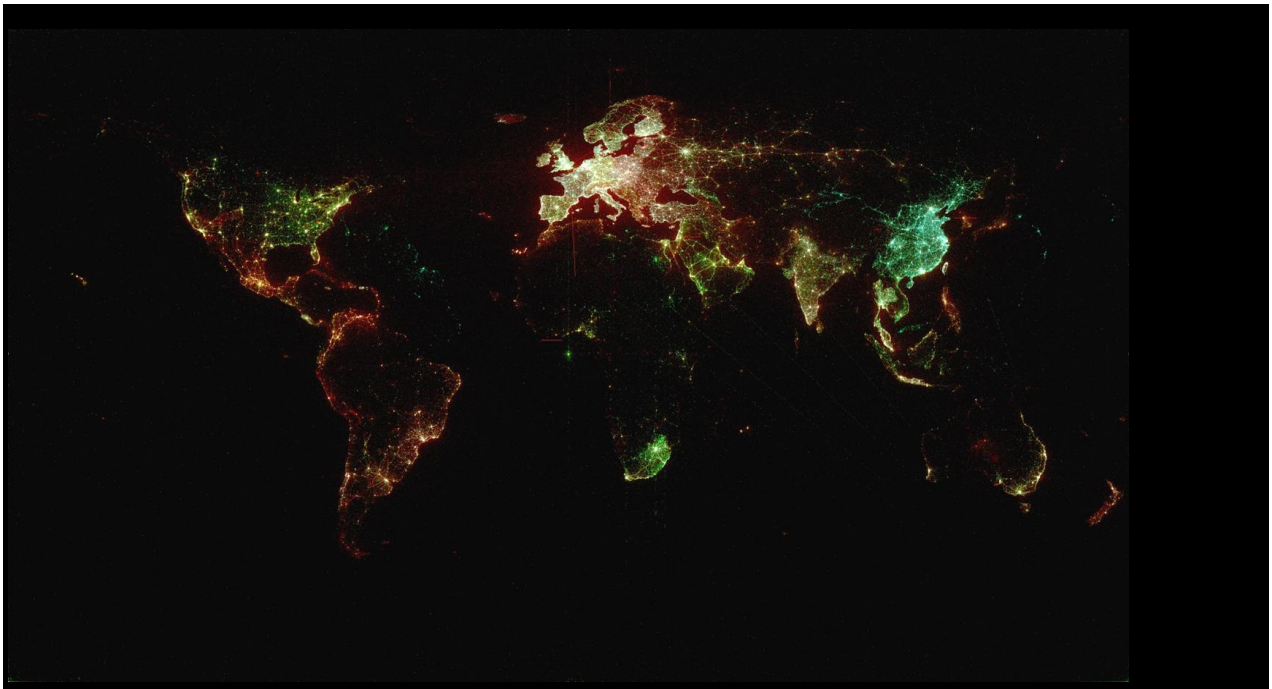
New Opportunity for Grand Challenge Research

"Despite their importance for urban planning, traffic forecasting and the spread of biological and mobile viruses, our understanding of the basic laws governing human motion **remains limited owing to the lack of tools** to monitor the time-resolved location of individuals" [Gonzales et al., Nature]

- Microscopic Sensors: Mobile devices become global-scale real-time personalized mobile sensing probes in the sensory swarm
- Macroscopic Science: Wireless clouds for live analytics of global mobile user data and in-situ social science research experiments

➢ "Large Hadron Collider" for Computer&Social Science Research
  - HUBRIS: Human Behavioral Research Infrastructure and Services

➢ From Big Data to Big Social Science Research Challenges
  - Human Behavior Modeling and Understanding (and Predicting?)




routemonkey
what is this?
00:05:20
In..
City
London
From
anywhere
To
anywhere
During..
Season
Days
Mon – Fri
Sat Sun


routemonkey
what is this?
13:49:10
In..
City
London
From
anywhere
To
anywhere
During..
Season
Days
Mon – Fri
Sat Sun

Grand Challenge Opportunity

Experimental Social Science

Our "Large Hadron Collider"

Computing Megatrends

- Leading-Edge Supercomputing
  - Current TOP100 supercomputers are Petascale (10^{15} FLOPS) systems
  - Challenges for next 7 years: push towards Exascale (10^{18} FLOPS) systems
  - Must improve performance/power efficiency from 1 GF/W to 100 GF/W
- Web Based Cloud Computing
  - Push towards cloud computing creates huge network bandwidth demands
  - Tension will result in federated and fragmented cloud computing models
  - Wireless edge of the cloud will be core to computing and communication
- Personal Mobile Computing
  - Continuation of Moore's law expected for at least two more process nodes
  - 100 GF/W technology can provide mobile supercomputers for mass market
  - Dealing with legacy SW and device installed base will be a huge challenge

Mobile Supercomputing

- Mobile Supercomputers
  - Improving performance/power efficiency to 100 GFLOPS/W will enable a Terascale (10^{12} FLOPS) mobile supercomputer with a 10W power budget.
  - An airborne supercomputer capable of 100 TFLOPS can then be deployed in an UAV (e.g. the RQ-1 and MQ-1 Predator drone) with a 1KW power budget.
- Architecture Innovations
  - Dataflow driven execution model supported by powerful SW tool chain and programmable and extremely energy-efficient HW fabric will be essential.
  - Exploitation of massive data level parallelism and inherent data redundancy.
- Form Factor Innovations
  - Extreme integration via 3D TSV die stacking of diverse technology dies, e.g. many-core processors, high-BW DRAMs and SSDs, FPGA, and power delivery.

Emerging Killer Applications

- Real-Time Environmental Sensing and Processing
  - Highly mobile and autonomous real-time data collection, data analytics, and data inference, without having to off-load to some remote cloud infrastructure.
  - Example: real-time traffic, special events monitoring, disaster management.
- Rapid Situational Deployment of Cloud Resource
  - On-demand localized wireless service can be provided by an UAV equipped with an airborne base station for special needs, e.g. a major sporting event.
  - Such drone based resource can provide very low latency and high bandwidth local services and function as the edge of the traditional cloud infrastructure.
- Swarm-of-Drones Infrastructure for Demanding Scenarios
  - Swarm of collaborating drones can be rapidly deployed to provide wireless communication and Petascale (10^{15} FLOPS) supercomputing infrastructure.