System Integration Implications for the IoT era (a tutorial)

Subu Iyer (s.s.iyer@ucla.edu)

chips.ucla.edu
Outline

• Macro trends
• Mobile, Data center and cognitive computing
• Semiconductor technology
• System Integration Technology
  – Packaging
  – 3D
  – Emerging
• Summary
Gartner Hype Cycle

Two years for IoT
Fundamentally, what is the IoT?

- Semiconductor ckt cost has come down linearly and are now very cheap.
- But Bandwidth has increased exponentially and also decreased in cost.

Network world (2012)
Fundamentally what is the IoT?

- Per capita BW is also increasing exponentially
- We now live in one huge interconnected system
- Most functions will be autonomous and with minimal human intervention
Fundamentally what is the IoT?

• The IoT started with the very first internet connection

• What we are seeing now is only a matter of degree NOT a fundamental paradigm shift

• But the scale of these interconnected devices
  – Depends on our ability to make them ubiquitous
  – The ability to sustain their interconnections
  – And transform data to information to knowledge to intelligence
Ubiquitous interconnected devices

Cost
Power size
Reliability
Fault tolerance
Configurability
The ability to sustain wireless and wired bandwidth growth

- Ability to handle higher data rates
  - Digital processing of analog signals – CMOS radios
  - In-door communication
  - Higher frequency operation
    - spectrum
  - Optical Fibers – will continue to be the backbone
  - Better coding techniques

- Broadcast Power management
  BW capability may be increasing faster than Moore’s Law
Transforming Data to Intelligence

• Big Data still has a big role to play
• Cloud platforms will still have to do the heavy lifting
• Cognitive computing will be critical
• Human intervention will be minimal
The Question

• Does the IoT mean that semiconductor electronics needs to change course?

Absolutely not!

The IoT era is a consequence of the progress in semiconductor technology – but this progress stalling

What can we do to help it along ...
Silicon Technology Drivers for IoT

**Mobile**
- More ultra low power heterogeneous integration
  - Sensors of all types
  - More analog
  - Mobile energy sources
  - Form factor
- Application space specific
  - Robots
  - Vehicles
  - Wearables / ingestibles
- Neuromorphic engine?
- Approx computation

**Stable**
- Single thread performance
- Thermal solutions
- Power distribution & delivery
- Massive Diverse Memory integration with small footprint
- Application specific processors eg. NMPs
- Advanced packaging / photonics

As bandwidth becomes “free”, mobile may be mostly data accumulation and aggregation while all computation will happen in the cloud – stable may be the bigger driver long term.

(from report of the IEEE Rebooting Computing 2 summit)
The Power of System on a Chip (SoC)

Transistor scaling has made this possible

1964 - Transistor
SLT module
6 transistors, 4 resistors

2014 – POWER8 Processor
22nm SOI eDRAM technology, 650mm²
12 cores and 96MB of on-chip memory
4.2 billion transistors

But as SoCs have gotten more complex

- NRE costs have skyrocketed
- Time to market has become huge
- Manufacturing costs have grown
- and yields have plummeted

Courtesy IBM
Computing is getting data centric and heterogeneous
“cognitive” computing – brawn Vs. brain

D. Modha (IBM) SyNAPSE

- Scaling the interconnects represents a formidable challenge
- Memory and logic will be far more intimately connected

Scale this up to:
- $8.5 \times 10^{10}$ neurons
- $5 \times 10^{14}$ synapses
“cognitive” computing – brawn Vs. brain

It is unlikely that we can just scale our selves out of the huge gap ($10^3$ to $10^4$) in power and interconnect density.
Lets get back to Moore’s law and scaling

Chip-Scaling can and will continue to at least the “7 nm” node

It's not device physics!

But economics will hold us back

i.e. How low can we go?
The FinFET is the best device we’ve ever made

- Better electrostatics
- Voltage scalability

But FinFETs on SOI are even better

- All of the above
- No junction leakage
- No punch thru’

But Scaled FINFETs are expensive
Two big device issues

• Intrinsic device Parasitics
  – Structurally many capacitances increase
• Contact resistance is an issue

Even so, the scaling of devices to 14nm and beyond does in fact provide incremental benefit in power and density.

Though, benefit relative to transistor cost is debatable.

The big elephants in the room are lithography and process complexity.
The Real Challenge is interconnect scaling

At lower levels of the hierarchy:

- Materials: Liner/barriers dominate
- Physics: Edge scattering
- Patterning: Restrictive geometries

Net: interconnect resistances growing much faster than the traditional 1/k
The Ferrari is great car

But tow a boat with it ??!!
At the high-end (data centers etc.)

- The scaling of devices to 14nm and beyond does in fact provide incremental benefit
  - In power and perhaps in density

- Though, benefit relative to transistor cost is debatable

- The big elephants in the room are lithography, process complexity, and interconnect performance
For edge devices - a different calculus

- Performance is not demanding
- Devices are off most of the time
- Power is dominated by leakage
- And most importantly cost and total power dominate
Dropping operating voltage

![Graph showing energy consumption at different Vdd values.]

SLVT, FBB=0.8V
TT, 25C

 Courtesy: S. Kengeri (Global Foundries)
PD and FD SOI

Back Biasing in ultra thin Box SOI
Allows for dynamic Vt control
Matching low total power performance to the application

- Fully Depleted SOI technology does allow for acceptable performance at \(~0.4\text{V}\) while minimizing total power
- The key is electrical control of \(V_t\) through back bias – can be software controlled
- There is a question of cost and Supply chain
ARM Cortex A7 Implementation – Initial results

22FDX is the First Technology to demonstrate 0.4V operation capability at >500Mhz on an ARM A7 Processor

- FinFet like Performance (1.2Ghz)
- 50% faster performance and 18% lower power than 28HKMG
- 47% lower power than 28HKMG at Iso-Frequency
- 92% Less Power at 520MHz (wrt 28HKMG at 800MHz)

22FDX at 0.4v

Source: Verisilicon

14nm FinFET

22FDX

50% Faster

+ 18% Less Power

47% Less Power

28SLP-HKMG

92% Less Power

Subi Kengeri (GLOBALFOUNDRIES)
Summary on device technology

• At the datacenter, its business as usual – more performance, lower active power, and a better cost performance tradeoff
  – The most advanced technology node helps as long as its within the cost constraints = patterning is the Achilles heel
  – Interconnect performance is the Achilles heel
• At the edge – its all about total power and where we can minimize standby power without a significant performance degrade at a low enough cost
  – The most advanced node is not ideal and a prior node with dynamically controllable threshold voltage is ideal
Putting the system together is much more than device technology

• Systems are very heterogeneous and diverse and that part of the problem has not garnered enough attention

• Going forward we cannot expect the customary improvements from scaling
What would Yogi Say?

So, if you can’t scale the chip
you should try to scale something else!
Package/Board Features have scaled modestly

Silicon features

1000X

Packaging features

~ 4X
A high performance board

Typical access latency in processor cycles (@ 4 GHz)

L1(SRAM) -> L2(SRAM/eDRAM) -> DRAM -> Flash -> HDD

High-Performance Disk
Consequences of Coarse Chip-Chip interconnects

- Chips are farther apart and sitting in an electrically lossy environment – increased latency, lower signal fidelity, impedance matching
- We can not put as many wires between chips
- So we need to employ serializers and deserializers (aka SerDes) – reduced bandwidth
- Links between chips need to transmit signals at high frequencies with fidelity – more area and power – gets relatively worse with scaling
Fig. 5: Approximate breakdown of power in a mid performance system. I/O power including memory power accounts for more than a third of total system active power.
Some other thoughts

Contrast the “app” environment with the “hardware” environment

- Time to market: few days Vs few years
- NRE cost Few $1000s vs a few $10′s of millions

- And our ability to integrate heterogeneous hardware is limited

- The IoT application space will be severely limited by these constraints
Do we need Packages?

Packages are supposed to:
- Protect the chip – mechanically and thermally
- Connect the chip electrically to other chips
- Allows mission mode testing of the chip

Epoxy /glass Board CTE = 16-20x10^{-6} /°C
Do we need Packages? – No!

Packages are supposed to:
- Protect the chip – mechanically and thermally (not really)
- Connect the chip electrically to other chips (by progressive fan-out)
- Allows mission mode testing of the chip (this is true)

We can scale the package by getting rid of it
And mount the bare die directly on the board

Epoxy /glass Board CTE = 16-20x10^{-6}\degree \text{C}
Evolution Of System Integration

“Prehistoric” Interposer/Boards
- Organic → Si → glass

Integration
- Stacked Die
- Stacked Memory

Now
- Wafer Stacking

Future
- Massively integrated Silicon-like Board

UCLA ENGINEERING
Henry Samueli School of Engineering and Applied Science
Birthplace of the Internet
Some Thoughts on 3D integration
Fig. 3: memory die stacked face-to-face using micro-bumps onto a processor die. All connections to the outside are made peripherally through wirebonds. This scheme works for low power chips.
Fig. 4 (a) A packaged multiple die 3-D stack (on an optional interposer) in the face-to-back embodiment. All connections have to go through TSVs. Typically all die are relatively low power.
Fig. 4 (b) A packaged two-high 3-die stack in the face-to-face embodiment. Only connections to the laminate go through TSVs. The top die is higher power and directly heat sunk.
Figure 5: Through-silicon vias (TSVs) may be introduced in a hierarchical wiring system. At lower levels, there is less wiring blockage but more integration challenges. Today’s relatively large TSVs are optimally integrated at wiring levels where dimensional discrepancies between the wiring level and the TSV are minimal. A couple (1–2) of wiring levels above the TSV can relieve a significant amount of wiring congestion.
Fig. 6 (a)
Courtesy IBM

Fig. 6 (a) shows a ceramic MCM. Multiple die are assembled on a multilayer ceramic substrate. Because Vias are mechanically punched, dimensions are rather large and this method of integration does not scale well. Going to organic laminates is only slightly better.

Fig. 6 (b)
Courtesy Semtech & IBM

Fig. 6 (b) shows the heterogeneous integration of SiGe BiCMOS die with a 45 nm SOI eDRAM ASIC on an interposer mounted on a ceramic substrate. On interposer, BW are in excess of 2TB/sec.
Fig. 7 (a) A HMC chip stack showing the lower logic die with four stacked memory die.
Courtesy Micron Technology And IBM

Fig. 7 (b) An interposer application with four stacked four-high HBM die and an AMD Graphics processor
Courtesy: AMD
Figure 9: Memory chip floor plan and power TSV overhead. See text for details.
Figure 10: 14nm embedded DRAM.

<table>
<thead>
<tr>
<th>Technology</th>
<th>14nm FinFET SOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Size</td>
<td>0.0174µm²</td>
</tr>
<tr>
<td>Macro Size</td>
<td>211.2µm x 284.8µm</td>
</tr>
<tr>
<td>Organization</td>
<td>4k x 144 x 2 Banks</td>
</tr>
<tr>
<td>Sub-Array</td>
<td>128 wordlines x 576 bitlines</td>
</tr>
<tr>
<td>Macro Performance</td>
<td>1ns Access / 2nS Cycle</td>
</tr>
</tbody>
</table>
Figure 11: Case study - C4, bump, TSV floorplan using 14nm embedded DRAM.
<table>
<thead>
<tr>
<th></th>
<th>DDR3</th>
<th>WIDE IO</th>
<th>HMC</th>
<th>HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference</strong></td>
<td>ISSCC 09</td>
<td>ISSCC 11</td>
<td>VLSI 12</td>
<td>ISSCC 14</td>
</tr>
<tr>
<td><strong>Target</strong></td>
<td>General</td>
<td>Mobile</td>
<td>General</td>
<td>General</td>
</tr>
<tr>
<td><strong>Node</strong></td>
<td>50nm</td>
<td>50nm</td>
<td>50nm</td>
<td>29nm</td>
</tr>
<tr>
<td><strong>Interface</strong></td>
<td>DDR3</td>
<td>CMOS</td>
<td>10G, 15G, (30G)</td>
<td>1 and 2Gb/s/pin</td>
</tr>
<tr>
<td><strong>Chip</strong></td>
<td>98 mm(^2) 2Gb</td>
<td>64mm(^2) 1Gb</td>
<td>68mm(^2) 1Gb</td>
<td>35mm(^2) 2Gb</td>
</tr>
<tr>
<td><strong>3D Module</strong></td>
<td>8Gb w/ 4 chips</td>
<td>2Gb/2 chips</td>
<td>4Gb w/ 4 chips</td>
<td>8Gb w/ 4 chips</td>
</tr>
<tr>
<td><strong>Peak Band</strong></td>
<td>12.8GB/s DIMM @ 800MHz</td>
<td>12.8GB/s chip @ 200MHz</td>
<td>128GB/s HMC @ 2GHz</td>
<td>128GB/s HMB @ 1GHz</td>
</tr>
<tr>
<td><strong>Min. TSV Pitch</strong></td>
<td>40(\mu)m x 60(\mu)m</td>
<td>40(\mu)m</td>
<td>~60(\mu)m</td>
<td>48(\mu)m x 55(\mu)m</td>
</tr>
<tr>
<td><strong>TSV R</strong></td>
<td>Info. NA.</td>
<td>0.25ohm</td>
<td>Info. NA</td>
<td>Info. NA</td>
</tr>
<tr>
<td><strong>TSV C</strong></td>
<td>300fF</td>
<td>47.4fF</td>
<td>Info. NA</td>
<td>Info. NA</td>
</tr>
<tr>
<td><strong># TSV</strong></td>
<td>300 Signals</td>
<td>&gt; 512</td>
<td>1886</td>
<td>&gt; 1024</td>
</tr>
</tbody>
</table>

Table 1
A unified road map for System Integration?
Shrinking the board

• Without question the smallest board is a single large die with all the system functions on it

  Amdahl eventually declared the idea would only work with a 99.99% yield, which wouldn't happen for 100 years.

• Hence the quest for the largest Yieldable Unit – the SoC (reticle size today will yield at 3-15%)
Packaging adopting wafer processing techniques

Wafer level Fan-out

eWLB = Embedded Wafer Level Ball grid array

Wafer level SiP

Courtesy nanium
Other developments

Silicon Bridge (Intel)

Intel Patent US 2014/0070380 A1

Apple watch

Chip Scale package

Flexible Electronics (PEL)

Silicon Photonics (IBM)
Mega SolfFs by re-integration on an Interconnect Fabric

The "right" interconnect fabric
- Mechanically robust (flat, stiff, tough...)
- Capable of fine wiring, fine pitch interconnects
- Thermally conductive
- Can have active and passive built-in components

Silicon Fits the Bill in many cases

Challenges:
- Warpage
- Topography
- Assembly/Thru’put
- Thermal

>100 μm pitch
Mass reflow

100μm > pitch > 40 μm
Mass re-flow + TCB

2-10 μm
Full contact – TCB
Proximate
Inductive
Capacitive
Flexible electronics

- Printed electronics has dominated mindshare

- But as heterogeneity and performance requirements become important assembly on flexible substrates will take share
Anatomy of a flexible sub-system

Power: source, storage, dissipation

Power distribution and control

Active layer: logic, memory

Sensing Layer: sensors, MEMS

I/O layer: RF, antenna, capacitive/inductive coupling

- Everywhere: flexible interconnects, wires, chip interconnects
- No individual die packaging – unconventional chip attach
Wafer to Wafer integration at tight pitch

4-layer wafer stack w/ 2.5 μm Pitch

DRAM Scaling also slowing
New Failure Modes

Highly interconnected Microsystems
3D Wafer Scale Integration Process

Make TSVs and the metal layer.

- Bond to handle.
- Remove handle wafer.

Thin the wafer a ~7μm processed layer on another.

- Aspect Ratio ~10
- Pitch < 5μm
- More than 40k/mm²

Wei Lin, et al., Prototype of Multi-Stacked Memory Wafers Using Low-Temperature Oxide Bonding and Ultra-Fine-Dimension Copper Through-Silicon Via Interconnects (2014)
Power vs Latency

- **2DI systems**
- **3DWSI systems**

Human in 2Di

Monkey in 2Di

Rat in 2Di

Human in 3Di

Monkey in 3Di

Rat in 3Di
CHIPS IoT Framework

Applications & Architecture
- Heterogeneous Systems
- Approximate Computing
- Cognitive Computing
- Fault Tolerance
- Supply chain Integrity
- Security
- Memory Subsystems
- Processing in Memory
- DFT
- Network on Board

New Tool Concepts
- Tool Development
- Scale-Up

Equipment

Materials
- Fine Pitch Interconnect
- Substrate Materials
- Warpage, Stress Flexible
- Materials
- Thermal solutions
- Materials for Additive Mfg.
- Reliability

Design infrastructure
- Thermo-Mechanical
- Electrical
- Tools
- Partitioning
- DFT
- Active IF Design

Integration
- A processing facility for Interconnect Fabric (IF) & assembly
  - *Silicon processing*
  - *Glass*
  - *Flexible substrates*
  - *Additive manufacturing*
  - *Thermal compression bonding*
  - *Wafer thinning*
  - *Wafer-wafer integration*

Devices/Components
- Novel switches
- New memory
- MEMS
- Sensors
- Passives, antennae
- Medical devices

Si, Compound Semis, MEMS, and OSATs

foundry
Data Fragmentation into dielets

SoC Chip Data

Foundry

Foundry 1

Foundry 2

Foundry

Foundry

Foundry

Foundry n

- Foundry source verification of dielets
- Reintegration on Si Interconnect Fabric
Chip Assembly
Summary

- The IoT is being hyped quite a bit – it's just a lot more of the same
- The key enablers have been:
  - Moore’s Law
  - Bandwidth growth
- What is missing is efficient ways to turn data into intelligence
  - This requires innovations in computing platforms – especially cognitive platforms
- The industry also needs an “app”-like hardware environment to stimulate innovation
  - The CHIPS approach has potential
Acknowledgements

I would like to thank my students and colleagues at UCLA and colleagues whose work I have quoted.
And Now for the Million Dollar Question

• Who will make money off the IoT?