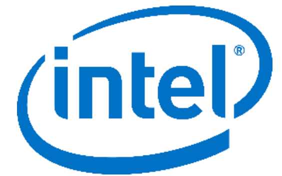
Memory Demand Trends and what they Mean to Packaging Technology

Ravi Mahajan

May 31, 2016

Key Contributors: Suresh Chittor, Randy Osborne, Bob Sankman



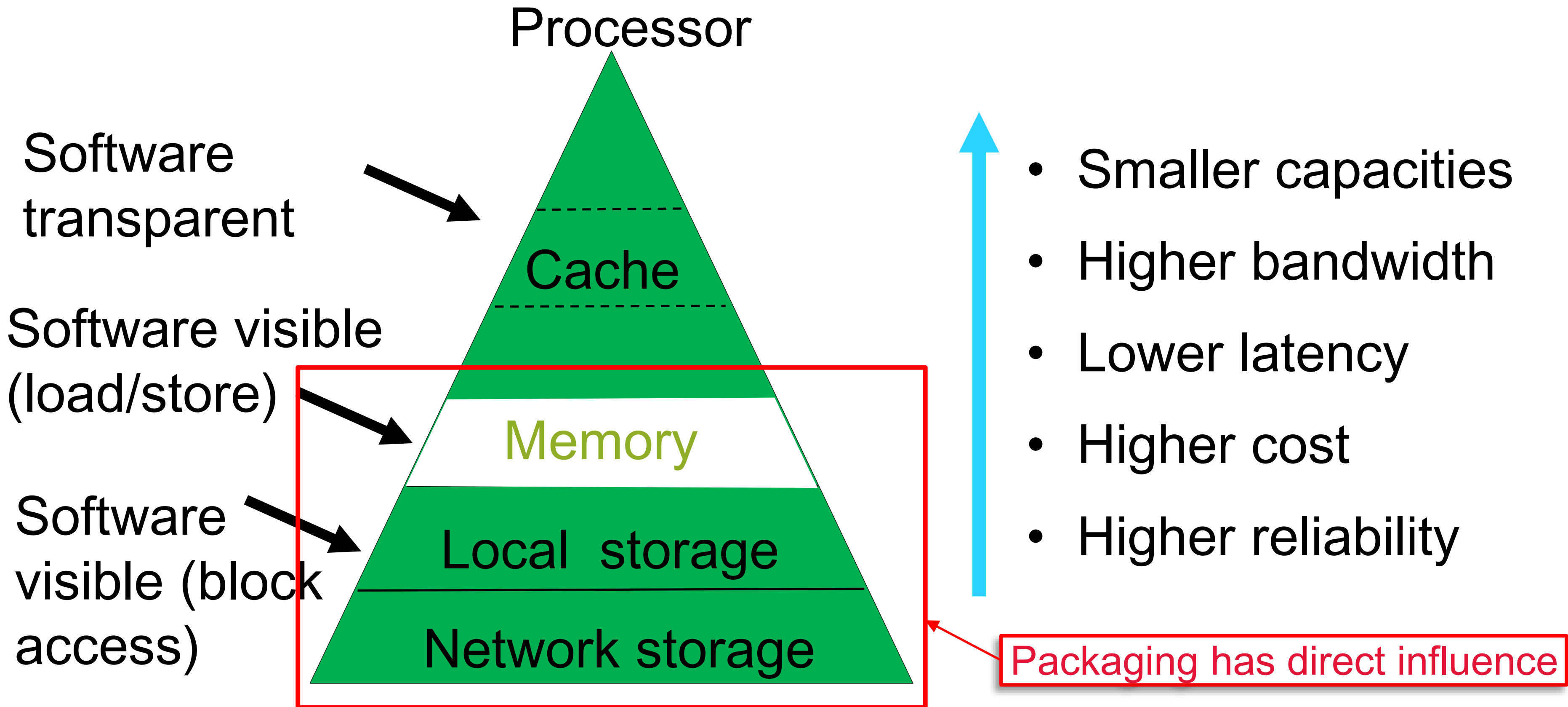
Key Messages

- CPU-Memory (DRAM) BW has increased > 2x every 2.5 years over the past decade (ISSCC 2016)
- Bringing Memory closer the CPU improves power efficiency & performance
  - High performance computing solutions already use On-Package memory to improve performance
  - Packaging innovations will continue to be needed to develop **cost effective MCP integration schemes** to support demand scaling and help proliferation of on-package memory integration

ECTC 2016 The 66th Electronic Components and Technology Conference

intel

Background: Memory Hierarchy (Suresh Chittor, ISSCC 2015)



Cache/Memory/Storage Hierarchy –
Key to optimizing performance/cost/power

Memory Hierarchy

Bandwidth

Higher power efficiency.

Standard cost, performance and power efficiency.

Lower cost.

Capacity

CPU

Far Memory Only

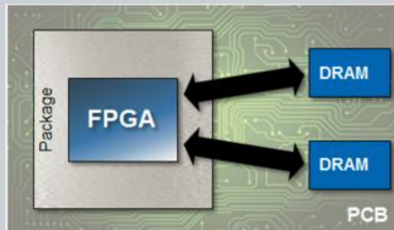
CPU

NM

FM

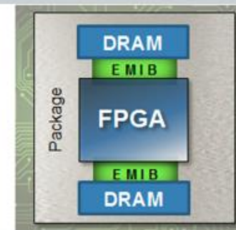
Near & Far Memory

On-Package Memory Integration Enables Higher Performance, Lower Power and Smaller Footprints



Far Memory

- Lower Bandwidth
- Higher Power
- Largest Footprint



Near Memory

- Highest Bandwidth
- Lowest Power
- Smallest Footprint

Near Memory vs. Far Memory

Altera's Stratix 10 DRAM SiP solutions are a near memory implementation, in that high-density DRAM is integrated very close to the FPGA, within the same package. In this configuration, the in-package memory is accessible significantly faster, up to 10X higher bandwidth, when compared to traditional main memory. A near-memory configuration also reduces system power by reducing traces between the FPGA and memory, while also reducing board area.

ECTC 2016 The 66th Electronic Components and Technology Conference

Speed & Data Width Trends

GPGPU : 4 x1024HBM

Servers – 4-8 x64 DDRx

Client - 4 x32 LPx/DDRx

Phones - 4 x16LPx

1000 GB/s

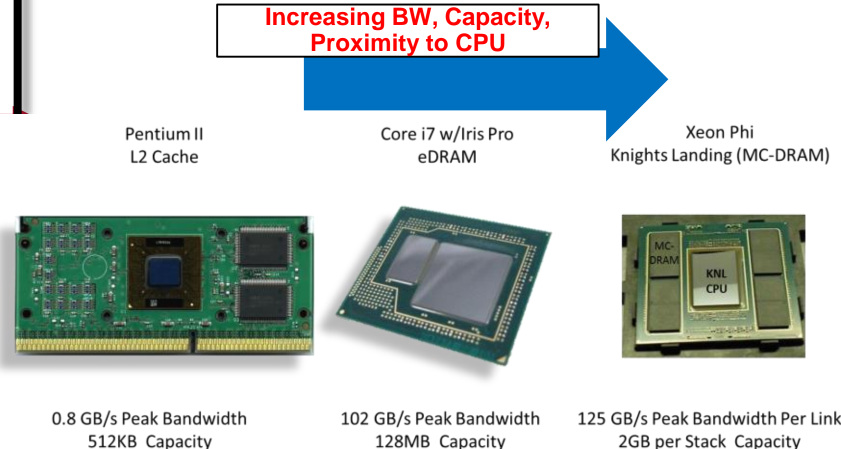
100 GB/s

Data Width/CPU: 4096, 2048, 1024, 512, 256, 192, 128, 64

Data Speed (GT/s): 800, 1600, 2400, 3200, 4000, 4800, 5600, 6400, 7200, 8000

*Data Width is total # of data lanes per CPU

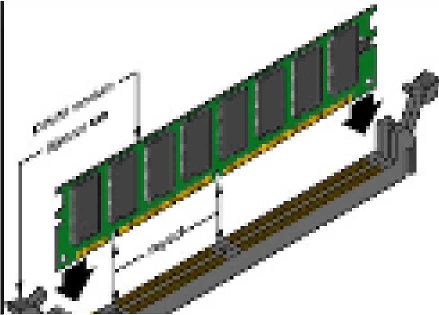
- Phones/Clients Systems push higher speeds with fixed data width for higher BW
- Servers use a mix of width and speed to for higher BW and capacity.
- GPGPU/GPU/Throughput computing moving to very wide channels with HBM, to support very high BW with very low power using on package interconnect



Memory : Industry trends

~1980 — 2015 — 2018+

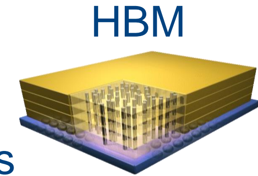
PCs, Servers



Server Desktop Mobile

RDIMMs
UDIMMs
LRDIMMs

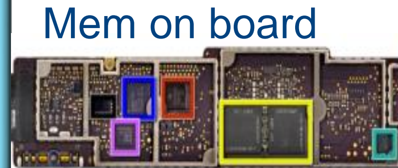
HBM



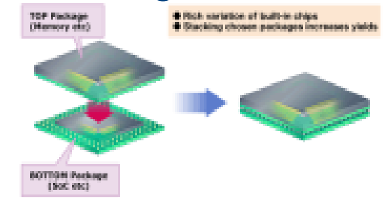
NVDIMMs, SCM

Tablets Phones

Mem on board



Package on Package

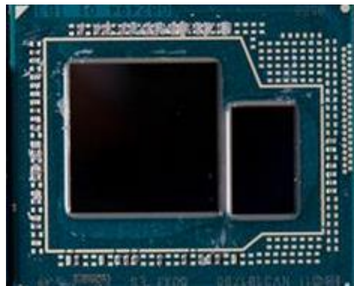


- Simple DIMM based Far Memory Solutions are evolving towards (Far + Near Memory) Solutions
  - Number of NVM solutions available or in development (See ISSCC 2016 Trends for a complete list)

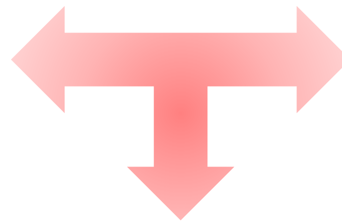
ECTC 2016 The 66th Electronic Components and Technology Conference

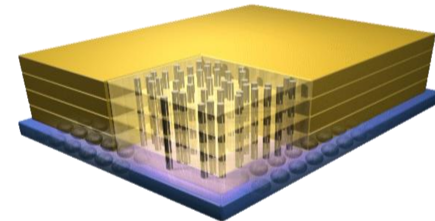
intel

Emerging devices and Form Factors



In-package
DRAM Cache





HBM - 3D Stacked DRAM

High power efficiency, but limited capacity per device.

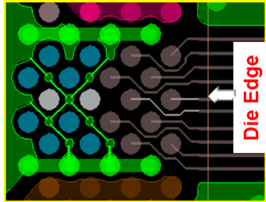
High BW and Low latency

Other High capacity DIMMs, lower cost, lower BW and higher latency in the works

Memory on board : More compact FF

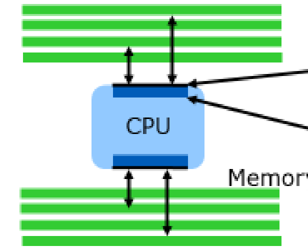
What Does this mean to Packaging?

Evolution of Dense MCPs



Key Package Design Metrics
- Wires/mm of Die Edge
- Signal Data Rate
- Energy/bit

CPU interface to memory.

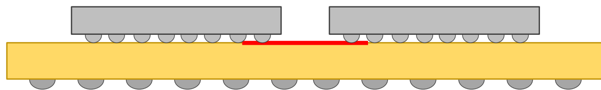


- IO perimeter (Si edge used) is the most challenged - needs to be optimized.
- IO area is also non-trivial, but more manageable than IO perimeter.

- Most CPU die are beginning to get IO limited with increased integration.
- BW per mm of die edge needs to continue to improve => Number of wires per mm, bits per wire, and speed need to improve.
- Need to optimize CPU IO for perimeter, area and power.

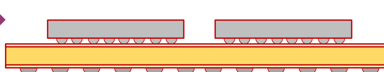
Suresh Chittor ISSCC Forum 2015 24

IO/mm/lyr = 28-34

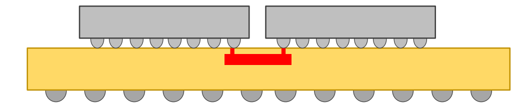


FCXGA,

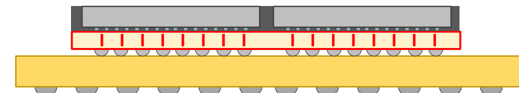
IO/mm/lyr 103*



HDI Organic Package/Interposer



EMIB



Silicon Interposer

IO/mm/lyr = 250

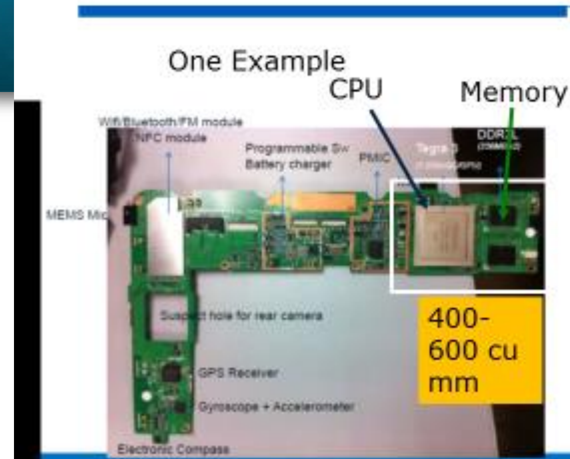
* Oi et. al. 2014 ECTC report 2µm L/S, 25µm pad

- Large number of IO's needed for Wide and Slow busses (e.g. HBM @ 1024 bits)
- Need power efficient, wide interconnects between CPU and memory

Packaging Challenges

- Power efficient on and off-package Memory Links
- Effective component and system Thermal solutions for the Entire CPU-Memory complex
- Cost effective interconnect scaling to for Dense MCPs and PoP solutions

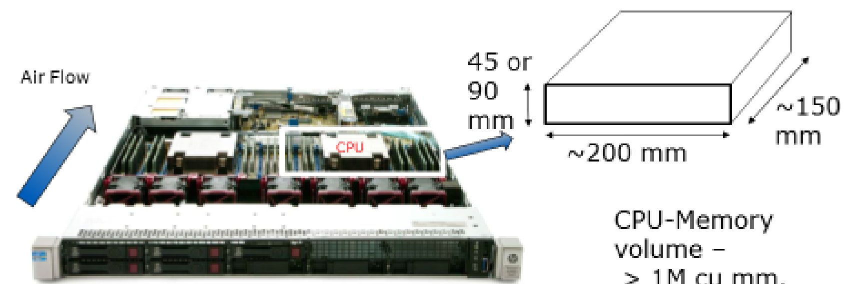
Form-factor : Phones/tablets



- Double sided board, X/Y and Z constrained.
- Total board area for electronics shrinking.
- CPU package is 200-300 sq mm, <1 mm thickness.
- Memory device and package size comparable to CPU, located adjacent (on side or top of) to CPU.

CPU and memory thermally coupled. Memory power affects CPU power and performance.

Form-factor : Servers



- Rack (shown above) or blade FF used for server platform. Typically 1U or 2U height for racks (1.75" or 3.5").
- Significantly larger volume for CPU+memory but also need many memory devices (compared to phones/tablets).
- Platforms are thermally constrained. Memory power limited.

Suresh Chittor ISSCC Forum 2015 23

ECTC 2016 The 66th Electronic Components and Technology Conference

intel

Key Messages

- CPU-Memory (DRAM) BW has increased > 2x every 2.5 years over the past decade (ISSCC 2016)

- Bringing Memory closer the CPU improves power efficiency & performance
  - High performance computing solutions already use On-Package memory to improve performance
  - Packaging innovations will continue to be needed to develop cost effective MCP integration schemes to support demand scaling and help proliferation of on-package memory integration

ECTC 2016
The 66th Electronic Components and Technology Conference

intel