Stallings 2012 edition

CS 320
Ch. 18 Multicore Computers

Multicore computer: **Combines two or more processors (cores) on a single die. Also called a chip-multiprocessor.**

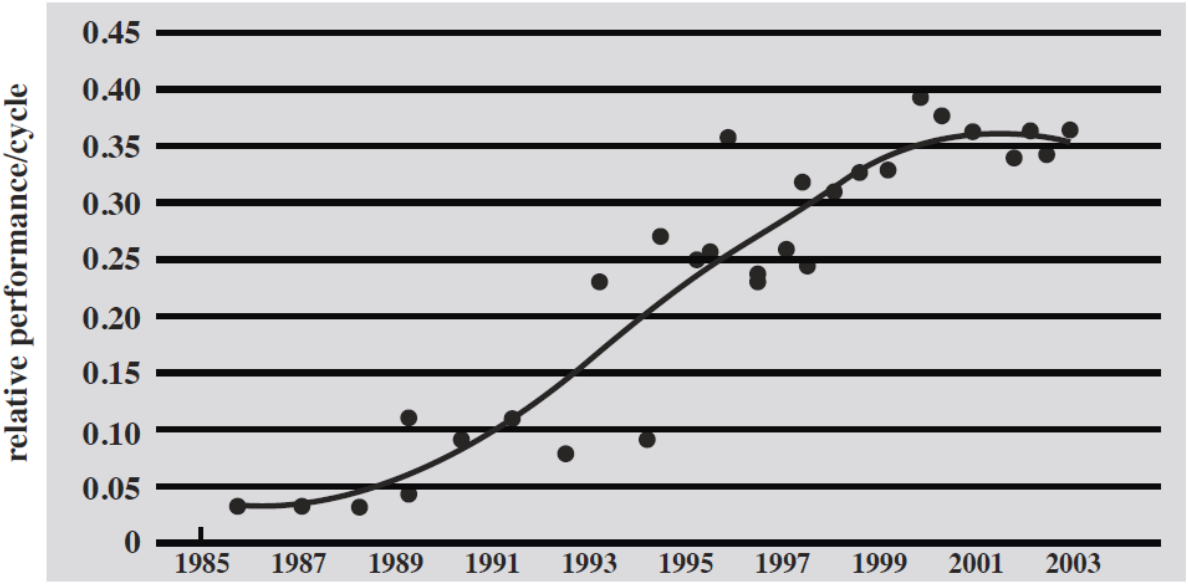
Definitions:
Hyper-threading – Intel's proprietary simultaneous multi-threading
Virtual core (aka logical core) – duplicates the register set in one core so that the execution unit can be used by two threads simultaneously.
Superscalar core – have multiple pipelines each with their own execution unit to allow parallel execution.
Simultaneous multithreading – the register banks are duplicated so that multiple threads can have their own register set but share execution resources in a pipeline.
Homogeneous multicore processor – a multicore processor with identical cores
Heterogeneous multicore processor – a multicore processor with different cores that may not share the same instruction set.

The architectural elements are in a typical core are:
Register set
ALU loop
Pipeline hardware
Control unit
L1 instruction and data cache
Many also have L2 and L3 caches.

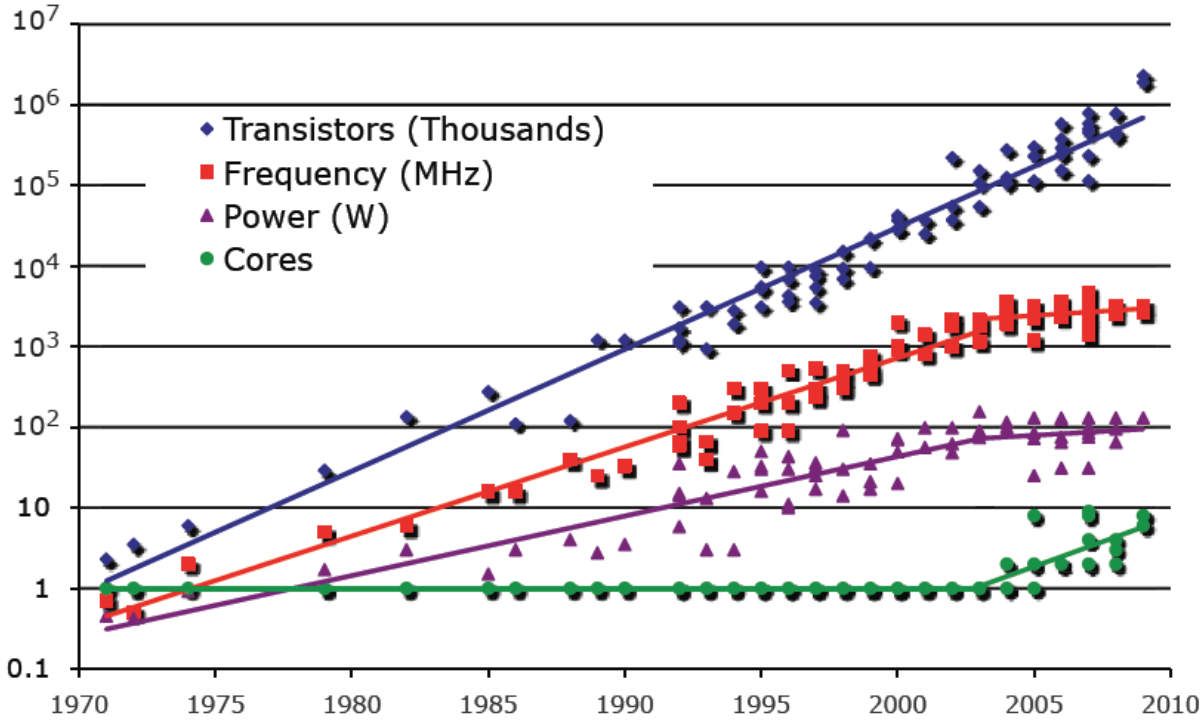
Prior to multicore computers there were three areas of architecture where parallelism was in use:
Pipelining
Superscalar architectures – multiple pipelines
Simultaneous Multithreading – Multiple register banks so that multiple threads can share the same pipe.

The figure below shows the relative performance/cycle vs time for Intel computers. **The flat region in 1985 gave way to an upward slope which was due to parallelism – mostly pipelining. A new flat region began in 2000 as we had gotten all that we could from pipelining and superscalar architectures.**

Stallings 2012 edition

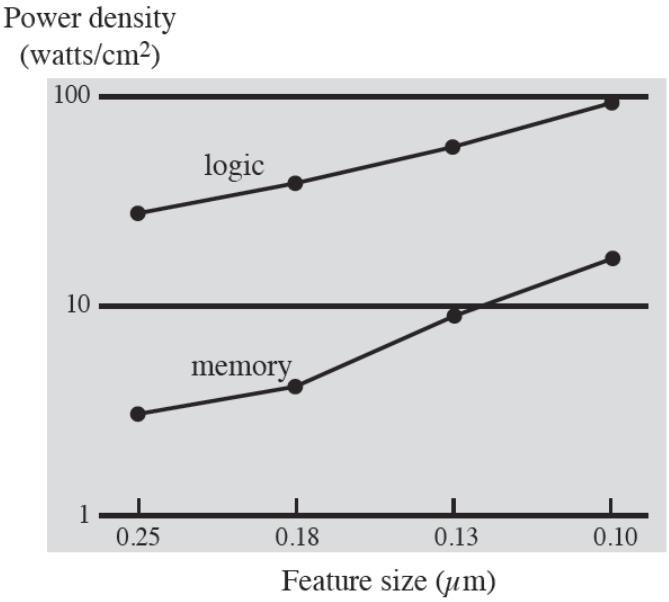


Power usage is also a serious problem as chip density increases. The figure below shows power usage for Intel porcessors. **The power has increased exponentially with time. In 2004/2005 the flattening of the power and frequency curves are due to multi-core chips.**



Stallings 2012 edition

There is a relationship between power density and whether the transistors on a CPU are used for memory or for logic. Memory transistors can be made much smaller than those used for logic and the power density of memory is an order of magnitude less than the power density for logic.

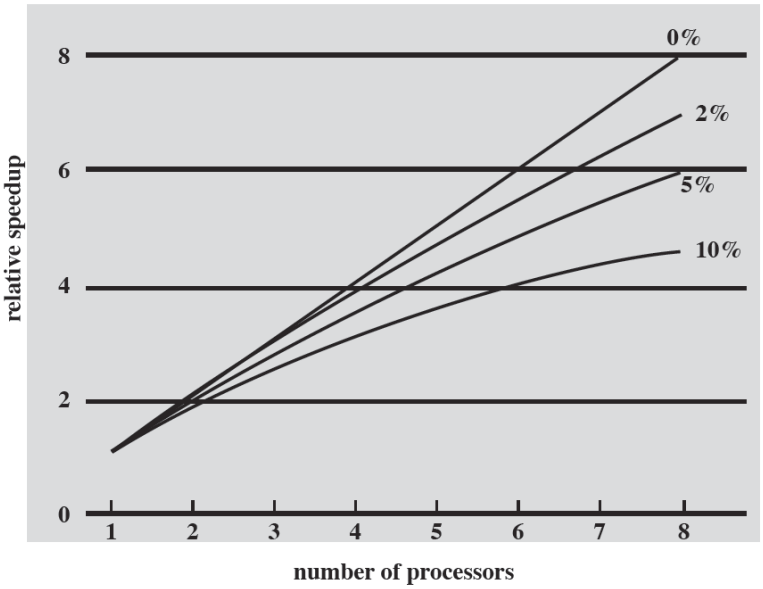


For a typical processor today about half of the chip area is used for memory or registers.

Pollack's rule states: **Performance increase is proportional to the square root of the complexity. This means that if double the logic in a chip you can expect about 40% more performance.**

Multicore computers have the potential to give a linear increase in performance vs complexity.

Software issues
"Multicore usage depends on the amount of serial code in an application." In the figure below, s**ome code is inherently serial in nature and cannot be run in parallel. Because of overhead and other factors, even a small amount (say 10%) of serial code can substantially reduce the effectiveness of multicore machines.**



Percent numbers are the amount of sequential code required.
This is from Amdahl's law.

$$Speed\ up = \frac{1}{(1-f) + f/N}$$

Where f = is the fraction of code which can be done in parallel and N is the number of processors.

The types of applications which can benefit from multicore computers are:

1) **Multi-threaded native applications**
2) **Multi-process applications**
3) **Java and C# applications. Provide scheduling and memory management for applications.**
4) **Multi-instance applications (multiple instances of the same program)**

Multicore Organization

There are many different multicore processor architectures, which vary in terms of:

Number of cores. Different multicore processors often have different numbers of cores. For example, a quad-core processor has four cores. The number of cores is usually a power of two.

Number of core types.

Homogeneous (symmetric) cores. All of the cores in a homogeneous multicore processor are of the same type; typically the core processing units are general-purpose **central processing units** that run a single multicore operating system.

Heterogeneous (asymmetric) cores. Heterogeneous multicore processors have a mix of core types that often run different operating systems and include **graphics processing units**.

Number and level of caches. Multicore processors vary in terms of their **instruction and data caches**, which are relatively small and fast pools of local memory.

How cores are interconnected. Multicore processors also vary in terms of their **bus** architectures.

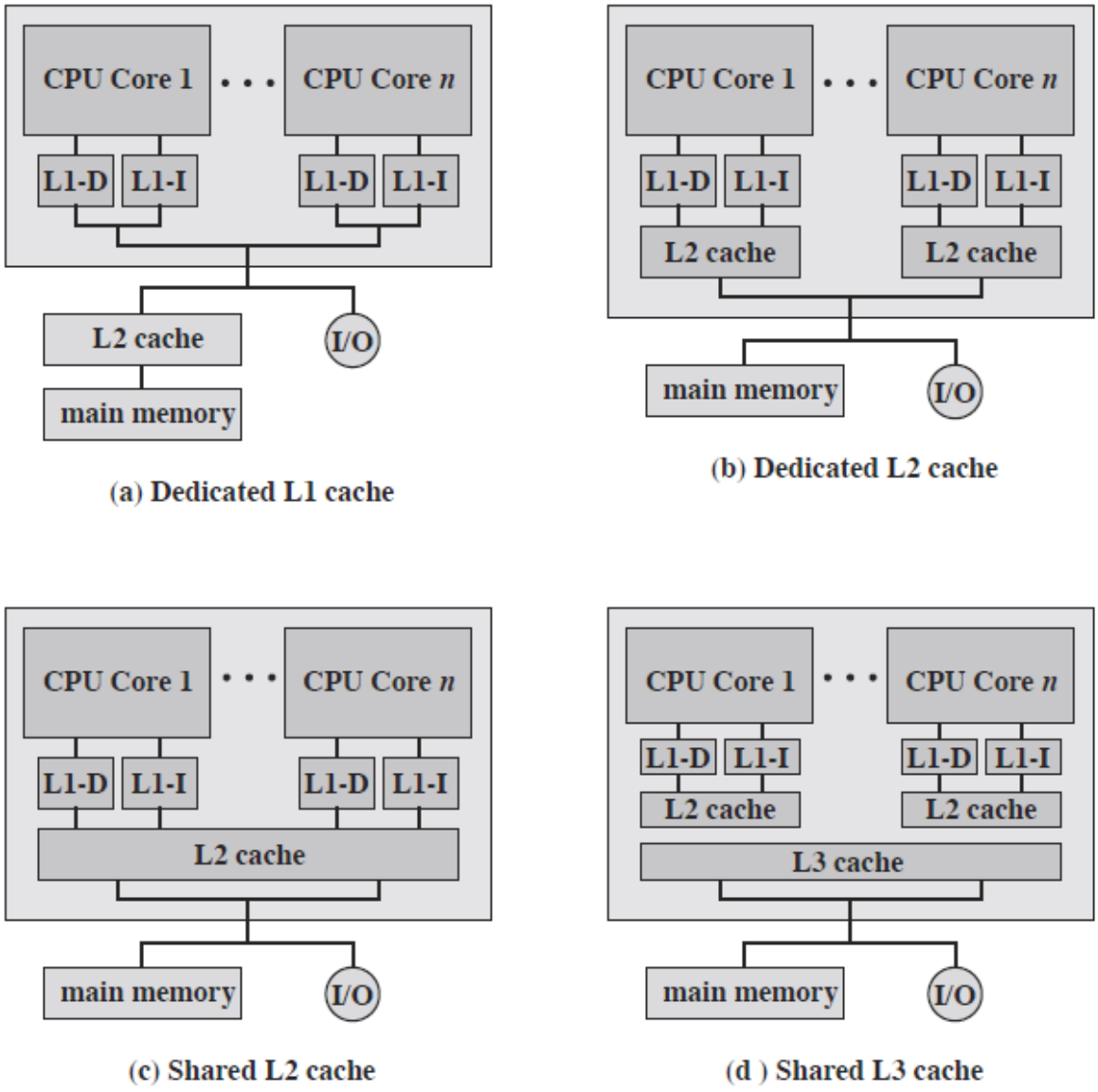
Isolation. The amount, typically minimal, of in-chip support for the spatial and temporal isolation of cores:

Physical isolation ensures that different cores cannot access the same physical hardware (e.g., memory locations such as caches and RAM).

Temporal isolation ensures that the execution of software on one *core* does not impact the temporal behavior of software running on another *core*.

Stallings 2012 edition

The figure below shows the different multicore architectures.



A) ARM 11 B) AMD Opteron C) Intel Core Duo C) Intel Core i7

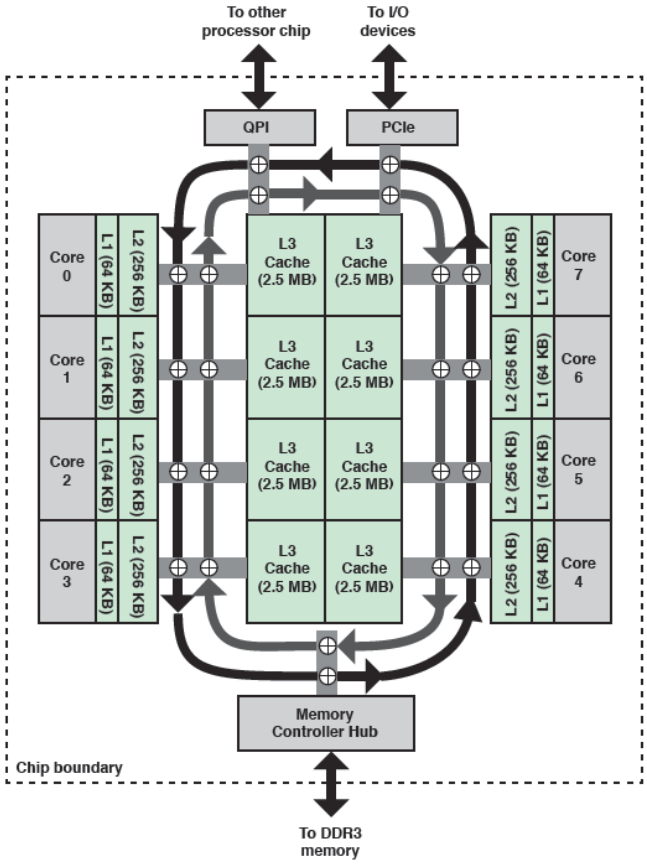
Note that the L1 **Data and instruction caches are almost always separate but the L2 and L3 caches are unified.**

The trend in multicore machines is to use a large shared L3 cache between cores and local L1 and L2 caches. This is preferred rather than having a large L2 local cache and no L3 cache.

Data and instructions brought into the L3 cache can be used by multiple cores if the cache is shared.

The shared main memory between cores also makes inter-process communication much simpler.

The figure below is that of the Intel Xeon E5-2600/4600 processor which illustrates the use of the large shared L3 cache.



Stallings 2012 edition

The figure below shows the general layout of Intel's Core-Duo processor.

- Thermal Control r**eports temperatures to software which alters clock speed.**
- Each core has an APIC. **This is an Advance programmable Interrupt Controller. Allows any core to interrupt another plus allows I/O interrupts and has timers.**
- Power management logic r**educes power consumption by lowering voltage levels when it can. Turns on individual units only when needed.**
- The 2 MB L2 cache is shared d**ynamically as needed.**
- **The Core Duo is a superscalar core.**

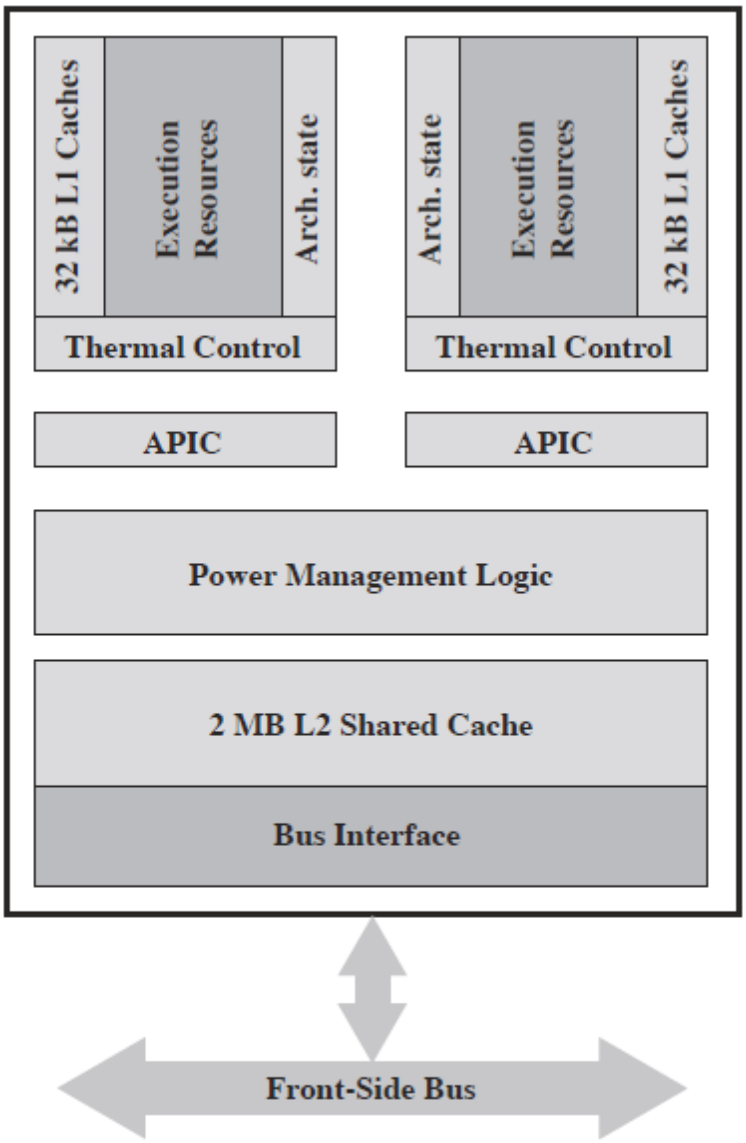
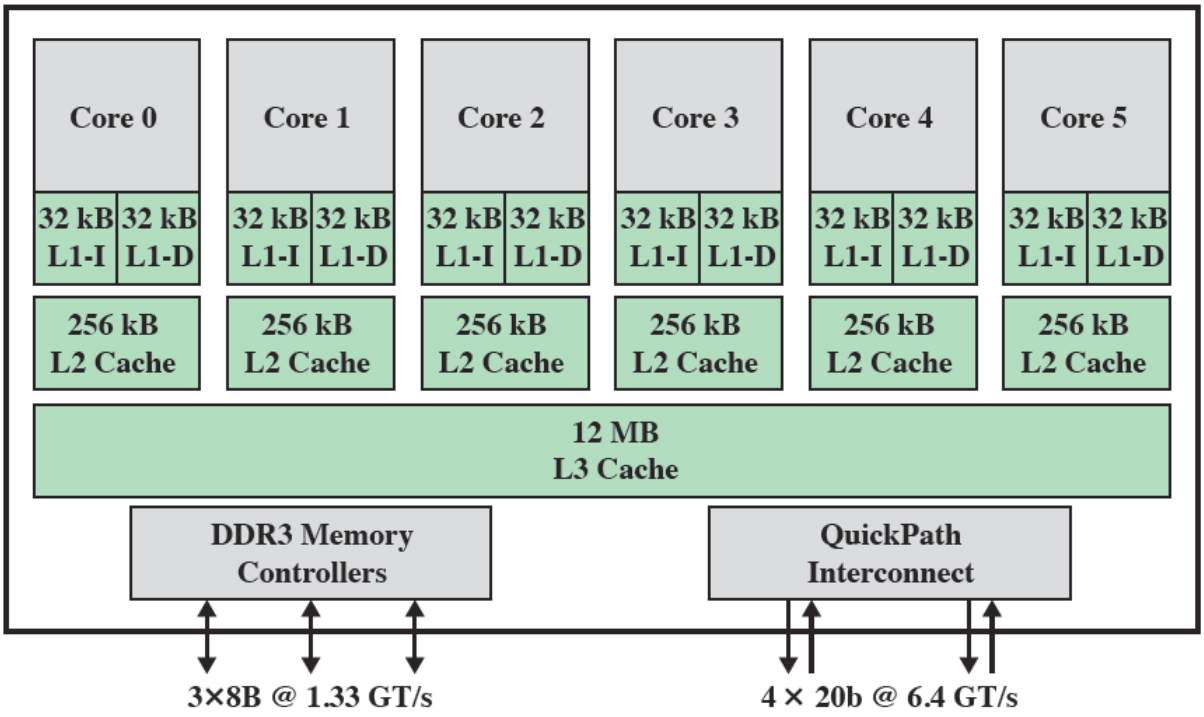


Figure 18.9 Intel Core Duo Block Diagram

Stallings 2012 edition

The figure below shows the Intel Core i7-990x processor.

- **The caches in this system use prefetching which speculatively loads the caches.**
- **There is a dedicated synchronous memory interface and a QuickPath connect that is a high speed port for communicating with other Intel cores.**
- **The i7 is an SMT core.**



All of the multi-core machines we have discussed to this point are homogeneous cores – that is, all of the cores use the same instruction set and the same hardware.

A heterogeneous multi-core processor is one in which the cores are not the same as each other and, in many cases, do not run the same instruction set.

The most common multicore heterogeneous machines combine a CPU with a Graphical Processing Unit or GPU.

Stallings 2012 edition

The figure below is that of the Texas Instrument 66AK2H12 which is a heterogeneous processor with 4 ARM Cortex-A15 and 8 C66x DSP cores on the same silicon.