19CSE310 GRID AND CLOUD COMPUTING

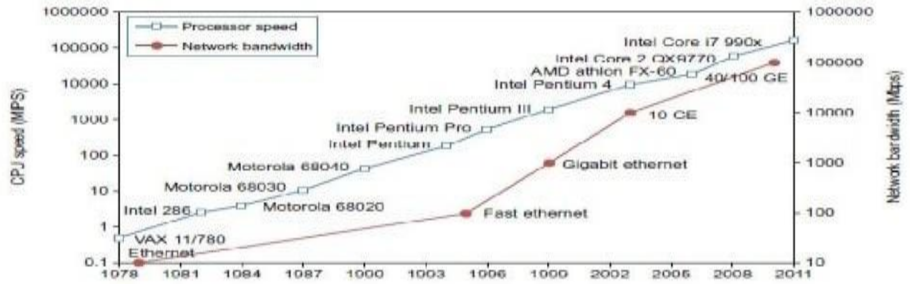
UNIT I INTRODUCTION

Evolution of Distributed computing: Scalable computing over the Internet – Technologies for network based systems – clusters of cooperative computers - Grid computing Infrastructures – cloud computing - service oriented architecture – Introduction to Grid Architecture and standards – Elements of Grid – Overview of Grid Architecture.

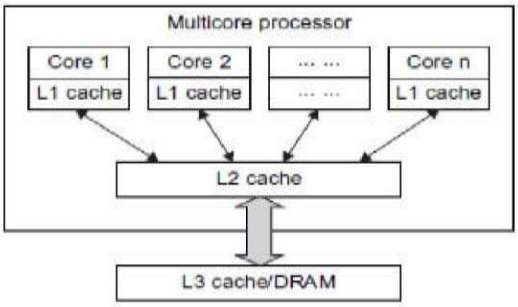
Technologies for network based systems:

1) Multicore CPUs and Multithreading Technologies

- A multicore architecture - with dual, quad, six, or more processing cores.
- These processors exploit parallelism at ILP and TLP levels.
- Processor speed growth is plotted in the upper curve in across generations of microprocessorsor CMPs.
- 1 MIPS - for the VAX 780 in 1978
- 1,800 MIPS - for the Intel Pentium 4 in 2002,
- 22,000 MIPS - the Sun Niagara 2 in 2008.
- As the figure shows, Moore law has proven to be pretty accurate in this case.
- The clock rate for these processors:10 MHz for the Intel 286,4 GHz for the Pentium 4
- However, the clock rate reached its limit on CMO S-based chips due to power limitations.
- The clock rate will not continue to improve unless chip technology matures.
- This limitation is attributed primarily to excessive heat generation with high frequency orhighvoltages.
- Both multi-core CPU and many-core GPU processors can handle multiple instruction threads atdifferent magnitudes today.
- Each core is essentially a processor with its own private cache (L1 cache).
- Multiple cores are housed in the same chip with an L2 cache that is shared by all cores. In the future, multiple CMPs could be built on the same CPU chip with even the L3 cache on the chip.
- Multicore and multi- threaded CPUs are equipped with many high-end processors, including the Intel i7, Xeon, AMD Opteron, Sun Niagara, IBM Power 6, and X cell processors. Each corecould be also multithreaded.



Improvement in processor and network technologies over 33 years.



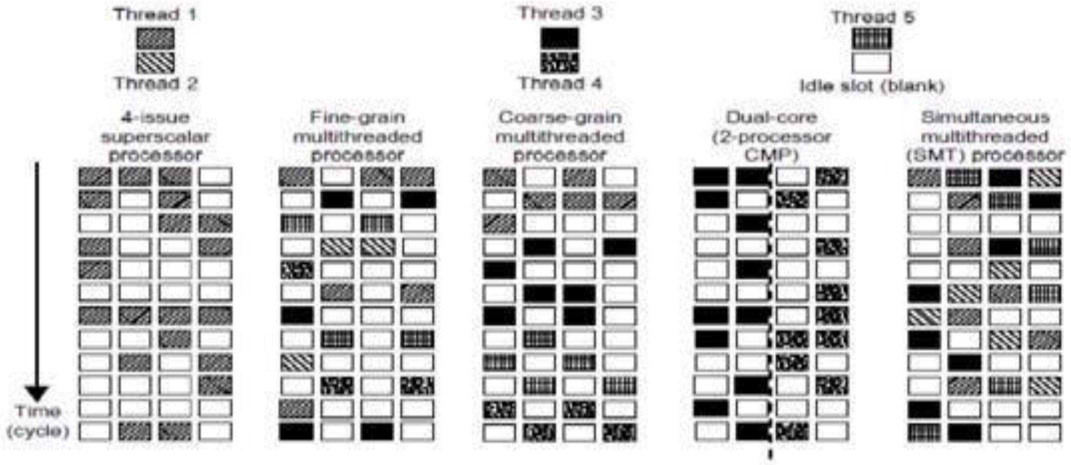
Schematic of a modern multicore CPU chip using a hierarchy of caches, where L1 cache is private to each core, on-chip L2 cache is shared and L3 cache or DRAM Is off the chip.

Multicore CPU and Many-Core GPU Architectures

- Multicore CPUs may increase from the tens of cores to hundreds or more in the future.
- But the CPU has reached its limit in terms of exploiting massive DLP due to the memory wallproblem.
- A many-core GPUs have with hundreds or more thin cores.
- Both IA-32 and IA-64 instruction set architectures are built into commercial CPUs. Now, x-86processors have been extended to serve HPC and HTC systems in some high-end server processors.

Multithreading Technology

The dispatch of five independent threads of instructions to four pipelined datapaths (functional units) in eachof the following five processor categories from left to right: a


Thread 1
Thread 2
Thread 3
Thread 4
Thread 5
Idle slot (blank)
4-issue superscalar processor
Fine-grain multithreaded processor
Coarse-grain multithreaded processor
Dual-core (2-processor CMP)
Simultaneous multithreaded (SMT) processor
Time (cycle)

Five micro-architectures in modern CPU processors, that exploit ILP and TLP supported by multicore and multithreading technologies.

The superscalar processor is single-threaded with four functional units. Each of the three multithreaded processors is four-way multithreaded over four functional data paths.

- **In the dual-core processor**, assume two processing cores, each a single-threaded two-way superscalar processor. Instructions from different threads are distinguished by specific shading patterns for instructions from five independent threads.
- **Fine-grain multithreading** switches the execution of instructions from different threads percycle.
- **Course-grain multi-threading** executes many instructions from the same thread for quite a fewcycles before switching to another thread.
- The multicore CMP executes instructions from different threads completely. These executionpatterns closely mimic an ordinary program.
- The blank squares correspond to no available instructions for an instruction data path at aparticular processor cycle.

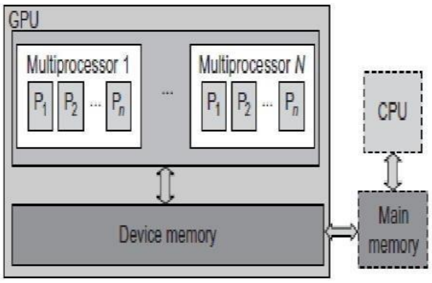
GPU computing to exascale and beyond

- A GPU is a graphics coprocessor or accelerator mounted on a computer's graphics card orvideo card.
- A GPU offloads the CPU from tedious graphics tasks in video editing applications.
- The world's first GPU, the GeForce 256, was marketed by NVIDIA in 1999.
- These GPU chips can process a minimum of 10 million polygons per second.
- Traditional CPUs are structured with only a few cores.
- For example, the Xeon X5670 CPU has six cores., a modern GPU chip can be built withhundreds of processing cores.
- GPUs have a throughput architecture that exploits massive parallelism by executing many concurrent threads slowly, instead of executing a single long thread in a conventional microprocessor very quickly.
- General-purpose computing on GPUs , known as GPGPUs, have

appeared in the HPC field.NVIDIA's CUDA model was for HPC using GPGPU.

How GPUs Work

- Early GPUs functioned as coprocessors attached to the CPU.
- Today, the NVIDIA GPU has been upgraded to 128 cores on a single chip.
- This translates to having up to 1 ,024 threads executed concurrently on a single GPU.
- Modern GPUs are not restricted to accelerated graphics or video coding.
- They are used in HPC systems to power supercomputers with massive parallelism atmulticore and multithreading levels.
- GPUs are designed to handle large numbers of floating-point operations in parallel.
- In a way, the GPU offloads the CPU from all data-intensive calculations.
- GPU are widely used in mobile phones, game consoles, embedded systems, PCs, and servers.
- The NVIDIA CUDA Tesla or Fermi is used in GPU clusters or in HPC systems for parallelprocessing of massive floating-pointing data.
- The interaction between a CPU and GPU for performing parallel execution of floating-pointoperations concurrently.
- The CPU is the conventional multicore processor with limited parallelism to exploit.
- The GPU has a many-core architecture that has hundreds of simple processing coresorganized as multiprocessors.



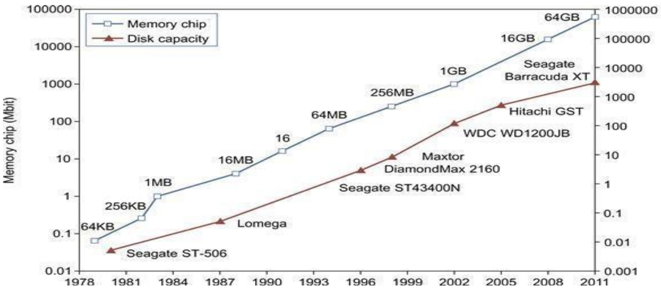
The use of a GPU along with a CPU for massively parallel execution in hundreds or thousands of processing cores.

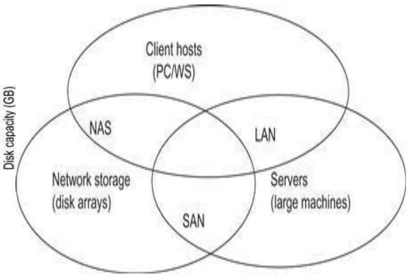
Power Efficiency of the GPU

- Bill Dally of Stanford University considers power and massive parallelism as the major benefitsof GPUs over CPUs for the future.
- By extrapolating current technology and computer architecture, it was estimated that 60Gflops/watt per core is needed to run an exaflops system
- Power constrains what we can put in a CPU or GPU chip.

MEMORY, STORAGE AND WIDE AREA NETWORKING

The upper curve in Figure 1.10 plots the growth of DRAM chip capacity from 16 KB in 1976 to 64 GB in 2011. This shows that memory chips have experienced a 4x increase in capacity every three years. Memory access time did not improve much in the past. In fact, the memory wall problem is getting worse as the processor gets faster. For hard drives, capacity increased from260 MB in 1981 to 250 GB in 2004.





System-Area Interconnects:

The nodes in small clusters are mostly interconnected by an Ethernet switch or a *local area network*(LAN). As Figure 1(B) shows, a LAN typically is used to connect client hosts to big servers. A *storage area network* (SAN) connects servers to network storage such as disk arrays. *Network attached storage*(NAS) connects client hosts directly to the disk arrays. All three types of networks often appear in a large cluster built with commercial network components. If no large distributed storage is shared, a small cluster could be built with a multiport Gigabit Ethernet switch plus copper cables to link the end machines.