INDIAN INSTITUTE OF TECHNOLOGY
ROORKEE

A SURVEY ON PROCESSOR SPEED UP TECHNOLOGIES

INTEL AND AMD

EC-252: COMPUTER ARCHITECTURE AND MICROPROCESSORS

UNDER THE GUIDANCE OF Dr. VASKAR RAYCHOUDHURY

SUBMITTED BY: Aman Kumar Rawat (10211004)

Anurag Chauhan (10211005)

Ashish Kumar Singh (10211006)

Intwala Nayan V. (10211014)

Pavan Patil (10214015)

Roshan Kumar S. (10211022)

Sudhir Kumar (10211025)

INDEX:

1. INTRODUCTION:

Computer performance is characterized by the amount of useful work accomplished by a computer system compared to the time and resources used.

Depending on the context, good computer performance may involve one or more of the following:

- Short response time for a given piece of work
- High throughput (rate of processing work)
- Low utilization of computing resource(s)
- High availability of the computing system or application
- Fast (or highly compact) data compression and decompression
- High bandwidth / short data transmission time

Processor being the heart and soul of the computer determines performance level of the computer. Since there are so many processors out there in the market, “WHAT MAKES THEM DIFFERENT?” The report here on focuses on various processors speedup technologies by the manufactures mainly the INTEL and AMD.

Beginning with the introduction to various technologies, we go through how they have been implemented in the architecture (INTEL being the major player). At the end to support the theoretical enhancements, a set of experimental data has been considered to be in congruence with the theory.

LEARNING OUTCOMES:

The report as whole is complete in itself. Readers should be able to able to get a holistic view of the topic.

Knowledge and understanding:

Through this report, emphasis has been laid on the study of microarchitecture and how the various technologies have been integrated in it. A significant increase in knowledge about the processors is promised.

Intellectual skills:

In the performance section, comparison between various processors has been shown. Any novice can easily compare processors with a given set of technologies to pick the best one.

Practical skills:

Knowing the intricacies of the technologies one can use the knowledge to boost performance of his/her PC to optimize the tasks like gaming.

2. PROCESSOR SPEEDUP TECHNOLOGIES OF INTEL

2.1) INTEL® TURBO BOOST TECHNOLOGY

INTRODUCTION:

Intel® Turbo Boost technology automatically allows processor cores to run faster than the base operating frequency if the processor is operating below rated power, temperature, and current specification limits. Intel® Turbo Boost technology can be engaged with any number of cores or logical processors enabled and active. This results in increased performance of both multi-threaded and single-threaded workloads.

C STATE DEFINITIONS:

Intel® processors based on *Nehalem microarchitecture* support core C0, C1, C3, and C6. C0 and C1 are always supported; the availability of the remaining C-states may vary by processor number. Any core within the processor can go into any C-state independent of the state of the other cores.

C0 State:
C0 is defined as the active state. While in C0, instructions are being executed by the core. For Intel® Turbo Boost technology, a core in C0 is considered an active core.

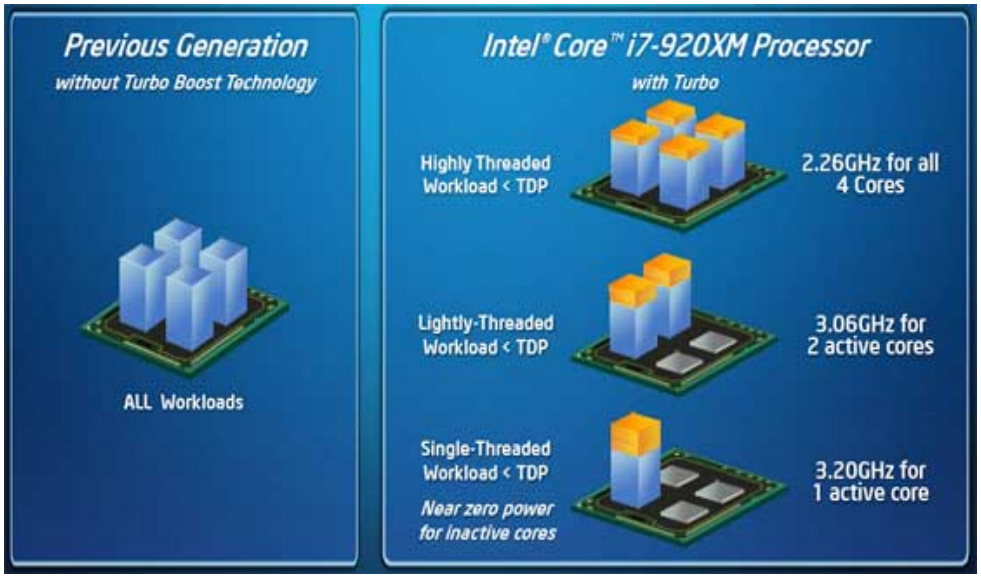
C1 State:
C1 is defined as the halt state. While in C1, no instructions are being executed. For Intel® Turbo Boost technology, a core in C1 is considered an active core. This idle state is generally classified as 'ACPI C1 '.

C3 State:
While in C3 the core PLLs are turned off, and all the core caches are flushed. For Intel® Turbo Boost technology, a core in C3 is considered an inactive core. This idle state may be classified as 'ACPI C2' or 'ACPI C3' depending on processor number, BIOS, and operating system.

C6 State:
While in C6, the core PLLs are turned off, the core caches are flushed and the core state is saved to the Last Level Cache. Power Gates are used to reduce power consumption to close to zero. For
Intel® Turbo Boost technology, a core in C6 is considered an inactive core.

HOW DOES IT ALL WORK?

The number of active cores at any given instant dictates the upper limit of Intel® Turbo Boost technology. For this discussion, a core is considered 'active' if it is in the C0 or C1 state; cores in the C3 or C6 state are considered 'inactive'. The upper limits will vary on a per processor number basis.
For example, one particular processor may allow up to two frequency steps (266.66 MHz) when just one core is active and one frequency step (133.33 MHz) when two or more cores are active. Therefore, higher deep C-state residency (C3 or C6) on some cores will generally result in increased core frequency on the active cores.



The upper limits are further constrained by temperature, power, and current. These constraints are managed as a simple closed-loop control system. If measured temperature, power and current are all below factory-configured limits and the OS is requesting P0, the processor automatically steps up core frequency (+133.33 MHz) until it reaches the upper limit dictated by the number of active cores.
When temperature, power or current exceed factory configured limits and you are above the base operating frequency, the processor automatically steps down core frequency (-133.33 MHz) in order to reduce temperature, power and current. The processor then monitors temperature, power, and current and continuously re-evaluates.

LIMITATIONS:

Intel® Turbo Boost technology core frequency upside availability is ultimately constrained by power delivery limits, but within those constraints, it is limited by the following factors:

- The estimated current consumption of the processor
- The estimated power consumption of the processor
- The temperature of the processor

Processors	Series	Max Clock speed	Threads	Unlocked Processors	Integrated Graphics
Core i7	2600K	3.4/3.8 GHz	8	Yes	HD3000
Core i7	2600	3.4/3.8 GHz	8	No	HD 2000
Core i5	2500K	3.3/3.7 GHz	4	Yes	HD 3000
Core i5	2500	3.3/3.7 GHz	4	No	HD 2000
Core i3	2XXX	3.3 GHz	4	No	HD 2000

OVERCLOCKING:

Overclocking is the technique used to increase the speed of a device beyond stock. 'Stock' is the name given to the default specifications of a device. Overclocking describes the alteration of clock speed above stock (hence the name overclocking) of any given device, usually RAM and CPUs, but also other components such as GPUs and PCI buses. When a device's clock speed is decreased from stock (or from an overclock), it is called Underclocking. The rule of overclocking states that "Maximize frequency and minimize voltage/current while removing as much heat as possible such that stability requirements are met."

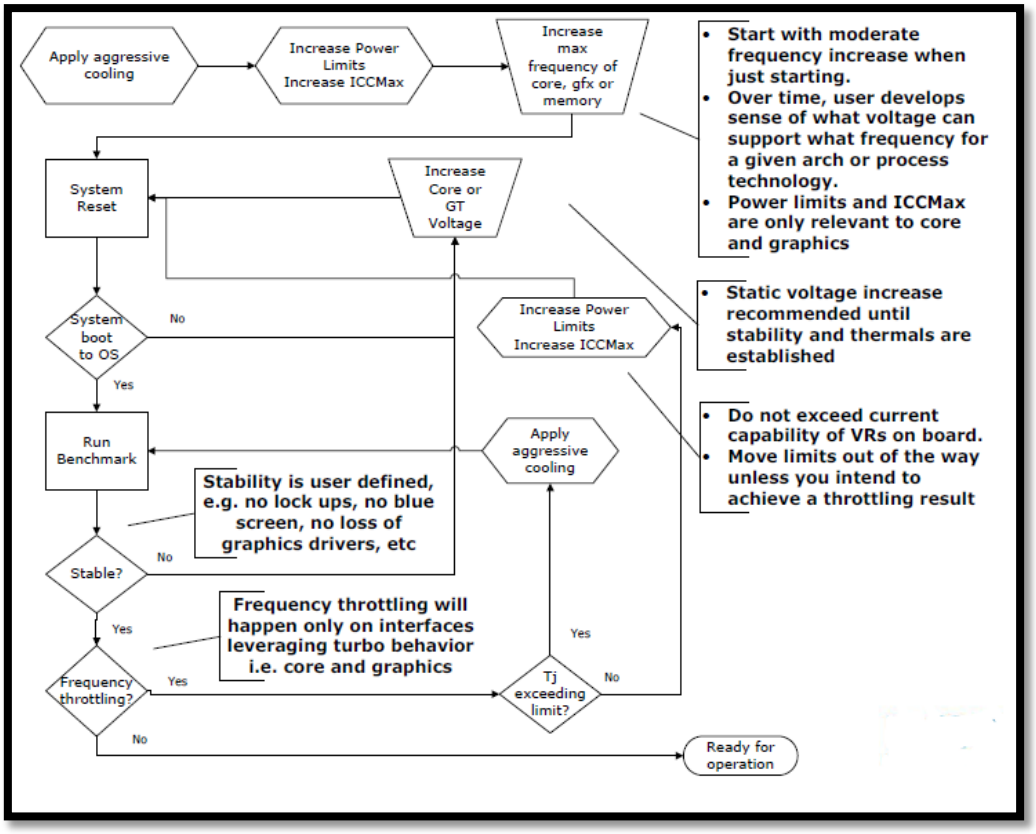
WHAT DETERMINES CPU PERFORMANCE?

The performance of a CPU is determined by its frequency (or clock speed), architecture and cache. The architecture of CPUs is constantly evolving and with this we get faster processors that run cooler and more efficiently. Multiple core CPUs are prime example of architecture improvement. With overclocking we can't change architecture because it is a physical component.

Clock speed is the rate in which data is processed. It is common to get confused with the idea and assume that a higher clock speed CPU runs faster than a lower clocked CPU. This is not necessarily true, especially when there is a major architecture difference.

Bottlenecking is the term used to describe a component in PC which is restricting other components due to lack of performance or efficiency. Overclocking can relieve a bottleneck if the problem lies with the CPU, RAM or GPU clock speeds.

OVERCLOCKING FLOW CHART:



2.2) INTEL HYPER-THREADING TECHNOLOGY (HTT)

INTRODUCTION:

Hyper-Threading Technology, a new microprocessor architecture technology. It makes a single processor look like two processors to the operating system. Intel's Hyper-Threading Technology delivers two logical processors that can execute different tasks simultaneously using shared hardware resources. Hyper-Threading Technology effectively looks like two processors on a chip. A chip with this technology will not equal the computing power of two processors; however, it will seem like two, as the performance boost is substantial. Chips enabled with Hyper-Threading Technology will also be cheaper than dual-processor computers: one heat sink, one fan, one cooling solution, and one chip are what are necessary.

Intel's Hyper-Threading Technology brings Simultaneous Multi-Threading to the Intel Architecture and makes a single physical processor appear as two logical processors with duplicated architecture state, but with shared physical execution resources. This allows two tasks (two threads from a single application or two separate applications) to execute in parallel, increasing processor utilization and reducing the performance impact of memory latency by overlapping the memory latency of one task with the execution of another. Hyper-Threading Technology-capable processors offer significant performance improvements for multi-threaded and multi-tasking workloads without sacrificing compatibility with existing software or single-threaded performance. Remarkably, Hyper-Threading Technology implements these improvements at a very low cost in power and processor die size.

WHAT LED TO THE ADVENT OF HYPERTHREADING TECHNOLOGY:

1. ILP (INSTRUCTION LEVEL PARALLISM) refers to techniques to increase the number of instructions executed each clock cycle. For example, a super-scalar processor has multiple parallel execution units that can process instructions simultaneously. With super-scalar execution, several instructions can be executed each clock cycle. However, with simple inorder execution, it is not enough to simply have multiple execution units. The challenge is to find enough instructions to execute.

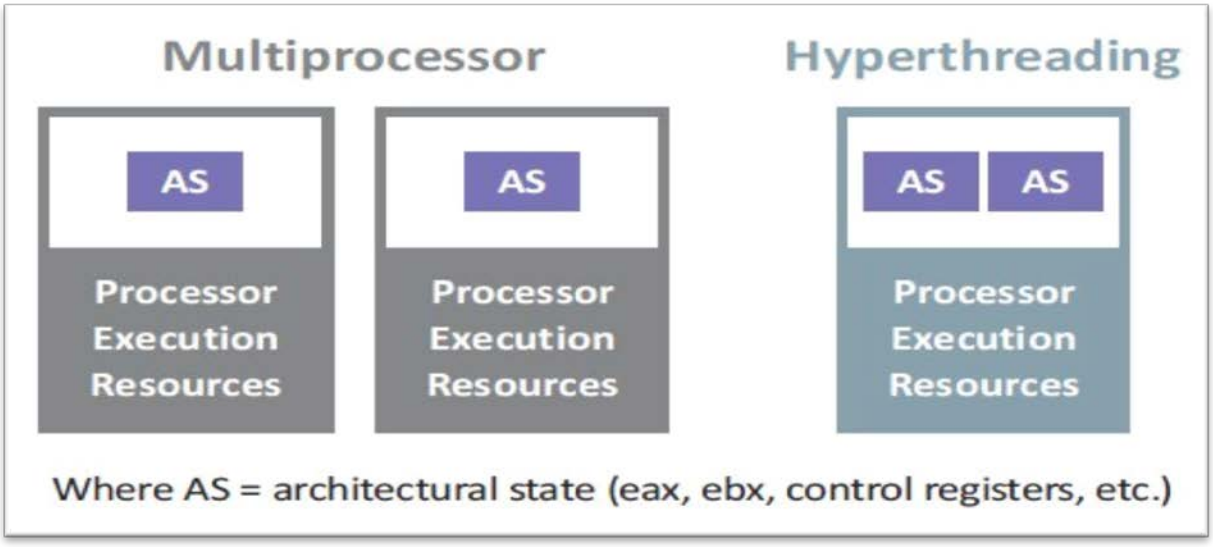
2. Accesses to DRAM memory are slow compared to execution speeds of the processor. One technique to reduce this latency is to add fast caches close to the processor. Caches can provide fast memory access to frequently accessed data or instructions. However, caches can only be fast when they are small. For this reason, processors often are designed with a cache hierarchy in which fast, small caches are located and operated at access latencies very close to that of the processor core, and progressively larger caches, which handle less frequently accessed data or instructions, are implemented with longer access latencies. However, there will always be times when the data needed will not be in any processor cache. Handling such cache misses requires accessing memory, and the processor is likely to quickly run out of instructions to execute before stalling on the cache miss.

3. The vast majority of techniques to improve processor performance from one generation to the next is complex and often adds significant die-size and power costs. These techniques increase performance but not with 100% efficiency; i.e., doubling the number of execution units in a processor does not double the performance of the processor, due to limited parallelism in instruction flows. Similarly, simply doubling the clock rate does not double the performance due to the number of processor cycles lost to branch misprediction.

HYPER-THREADING TECHNOLOGY ARCHITECTURE:

Hyper-Threading Technology makes a single physical processor appear as multiple logical processors. To do this, there is one copy of the architecture state for each logical processor, and the logical processors share a single set of physical execution resources. From a software or architecture perspective, this means operating systems and user programs can schedule processes or threads to logical processors as they would on conventional physical processors in a multiprocessor system. From a micro architecture perspective, this means that instructions from logical processors will persist and execute simultaneously on shared execution resources.

The figure above shows that, each logical processor maintains a complete set of the architecture state. The architecture state consists of registers including the general-purpose registers, the control registers, the advanced programmable interrupt controller (APIC) registers, and some machine state registers. From a software perspective, once the architecture state is duplicated, the processor appears to be two processors. The number of transistors to store the architecture state is an extremely small fraction of the total. Logical processors share nearly all other resources on the physical processor, such as caches, execution units, branch predictors, control logic, and buses. Each logical processor has its own interrupt controller or APIC. Interrupts sent to a specific logical processor are handled only by that logical processor.



FRONT END OPTIMISATION:

The front end of the pipeline is responsible for delivering instructions to the later pipe stages. As shown in Figure 5a, instructions generally come from the Execution Trace Cache (TC), which is the primary or Level 1 (L1) instruction cache. Figure 5b shows that only when there is a TC miss does the machine fetch and decode instructions from the integrated Level 2 (L2) cache. Near the TC is the Microcode ROM, which stores decoded instructions for the longer and more complex IA-32 instructions.

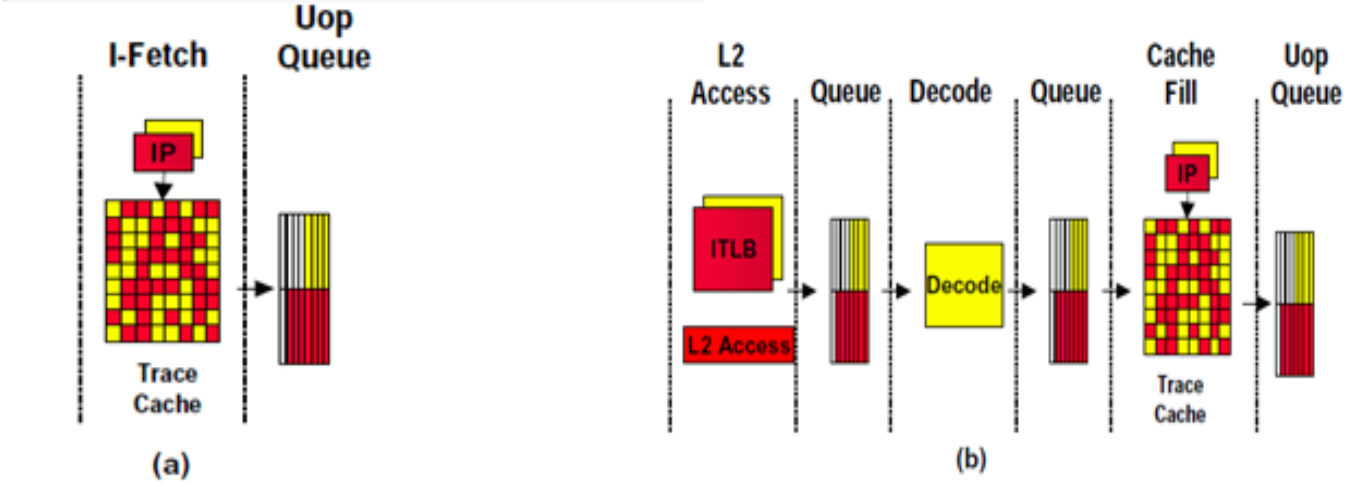


Figure: Front-end detailed pipeline (a) Trace Cache Hit (b) Trace Cache Miss

2.3) ENHANCED INTEL® SPEEDSTEP® TECHNOLOGY

INTRODUCTION:

The Intel SpeedStep Technology (EIST or SpeedStep short for "Enhanced Intel Speedstep Technology") is an energy saving feature in notebook and desktop processors from Intel .

Enhanced Intel® SpeedStep® Technology has revolutionized thermal and power management by giving application software greater control over the processor`s operating frequency and input voltage. Systems can easily manage power consumption dynamically.

NEED FOR SPEEDSTEP TECHNOLOGY:

Today`s embedded systems are demanding greater performance at equivalent levels of power consumption. Legacy hardware support for backplanes, board sizes and thermal solutions have forced design teams to place greater emphasis on power and thermal budgets. Intel has extended architectural innovation for saving power by implementing new features such as Enhanced Intel SpeedStep Technology.

The Enhanced Intel SpeedStep technology attempts to address the following challenges:

- Minimizing system and processor unavailability :

  Operating point switching requires voltage to be transitioned over a wide range (e.g., from 0.9V to 1.5V). Physical limitations of the power delivery system translate this demand to over 100µs delay. A full clock generator Phase-Locked-Loop relock requires approximately 30µs. The architecture needs to ensure system memory access unavailability will not exceed 10-15µs, to match isochronous device needs.

- Self-managed voltage and frequency stepping :

  The Enhanced Intel SpeedStep technology requires the migration of the mechanism from the chipset into the processor. This introduces two challenges: (a) how to sequence the operation when the processor clock is halted and (b) how to prevent loss of system events, such as interrupts and snoops, previously blocked by the chipset during the transition.

DESIGN & ARCHITECTURE STRATEGIES FOR SPEEDSTEP TECHNOLOGY:

Enhanced Intel SpeedStep Technology builds upon that architecture using design strategies that include the following:

- Separation between Voltage and Frequency Changes. By stepping voltage up and down in small increments separately from frequency changes, the processor is able to reduce periods of system unavailability (which occur during frequency change). Thus, the system is able to transition between voltage and frequency states more often, providing improved power/performance balance.
- Clock Partitioning and Recovery. The bus clock continues running during state transition, even when the core clock and Phase-Locked Loop are stopped, which allows logic to remain active. The core clock is also able to restart far more quickly under Enhanced Intel SpeedStep Technology than under previous architectures.

HOW DOES IT WORK?

Intel thermal and power management technology that allows the processor performance and power consumption levels to be modified while a system is running. Enhanced Intel SpeedStep Technology works via application software, which changes the bus-to-core frequency ratio and the processor core voltage. The operating state is determined by things such as the system power source, the processor thermal state, and even the operating system policy.

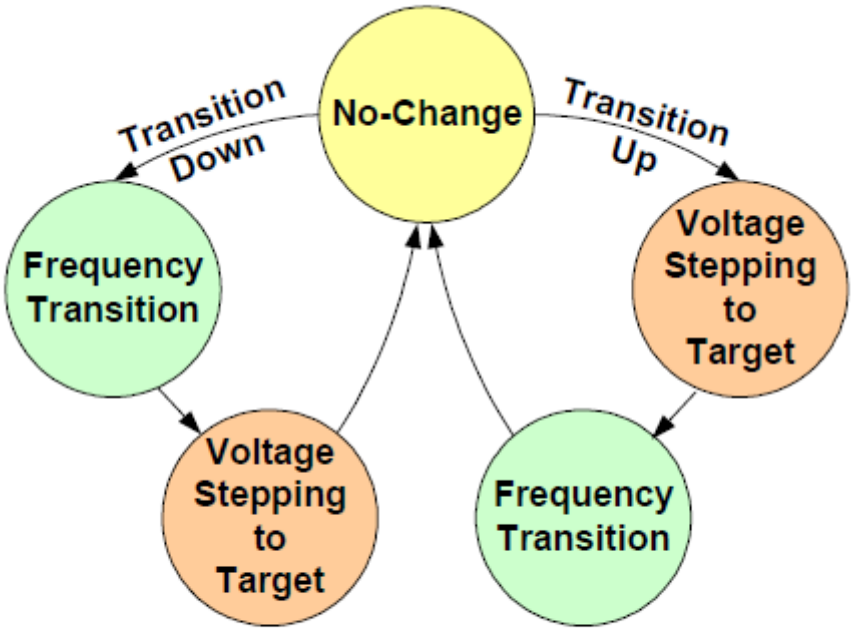
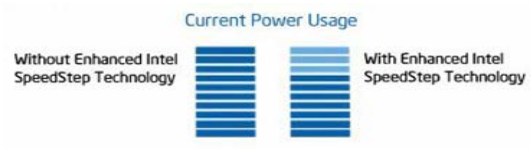
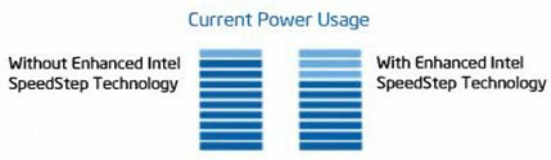


Figure : Enhanced Intel SpeedStep® technology transition sequencing

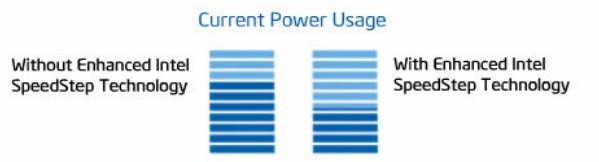
1. When processing workload is demanding EIST increase the power to meet the performance needs.



2. As the workload decreases EIST reduces the power to increase battery life and reduces fan noise also.



3. When workload is minimal EIST can further reduce power.



2.4) Intel® Smart Cache Technology

- **Intel® Smart Cache Technology (in** Core Microarchitecture and P4)
    i. Separate L1 cache per each core
    ii. Shared L2 cache among all cores instead of separate
    iii. L2 cache is non-inclusive
- **Intel® Advanced Smart Cache Technology (in** Nehalem and Sandy Bridge Microarchitecture)
    i. Separate L1 and L2 cache per each core
    ii. Shared L3 cache among all cores and Integrated GPU
    iii. L3 cache is inclusive

How it works?

1. Separate L1 cache provides for the data and instruction to processor very quickly due to low latency due to very small size and is very close to processor core.
2. Last level cache is shared across all the unit has two main advantage :
    - Increase in cache hit as all the unit have access to full cache not limited by partition
    - Full utilization of cache as all the units access it.
3. If cache is non- inclusive it takes snooping time to find if the data requested by the one core if not present in next level cache, is it present in cache of other cores? So cache- misses increase execution time.
4. If the cache is inclusive then if data requested by one core ,if it is not present in next level cache then it is implied that the data is not present in any of the cache of other cores. This reduces snooping time and increases performance by executing more instructions.

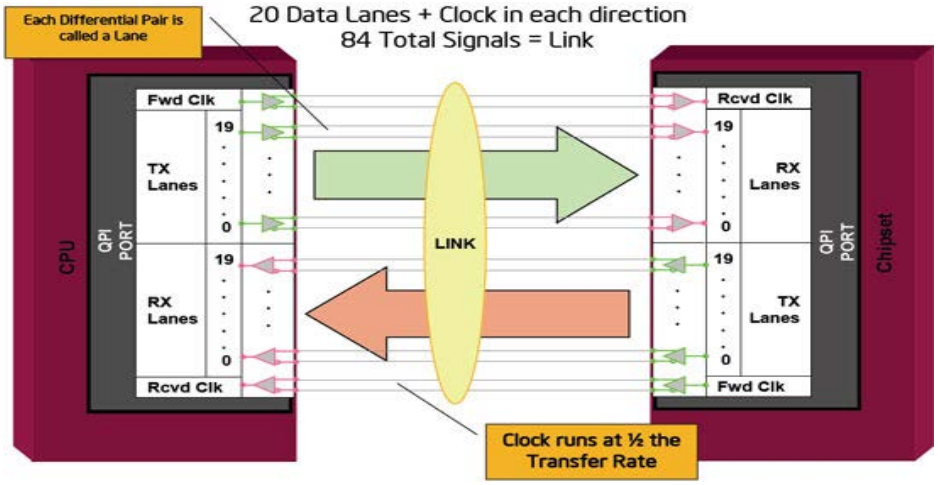
Difficulties in implementation :

- Cache bank conflict (core 2): Can't perform read, write on two location having same bank no. For e.g.,
    - mov eax, [esi] ; Use bank 0, assuming esi is divisible by 40H
    - mov [esi+100H], ebx ; Use bank 0. Cache bank conflict
    - mov [esi+110H], ebx ; Use bank 1. No cache bank conflict
- Miss aligned memory access (core 2, AMD-k10)
    - The Core2 has a penalty for misaligned memory access when a cache line boundary (64 bytes) is crossed.
- False memory location dependency (core 2, Nehalem, Sandy bridge) :have a false dependence between memory addresses with the same set and offset, i.e. with a distance that is a multiple of 4 kB.
    - mov [rsi], eax
    - mov ebx, [rsi+1000H] ; False memory dependence

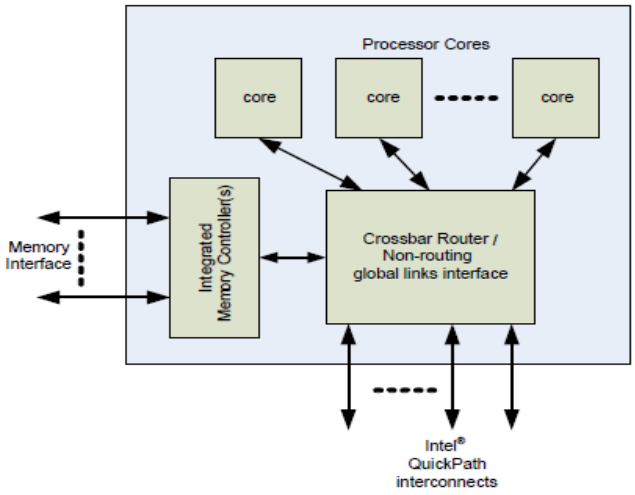
Cache	Core 2	Nehalem	Sandy Bridge
Level 1 code	**latency 3**	**latency 4**	**latency 4**
Level 1 data	**latency 3**	**latency 4**	**latency 4**
Level 2	**latency 15**	**latency 11**	**latency 12**
Level 3	**-**	**latency 38**	**latency 20**

2.5) Intel® QuickPath Technology

- The Intel® QuickPath Interconnect is a high speed packetized, point-to-point interconnect.
- The narrow high-speed links stitch together processors in distributed shared memory1-styleplatform architecture. Compared with today's wide front-side buses, it offers much higher bandwidth with low latency.
- It has a snoop protocol optimized for low latency and high scalability, as well as packet and lane structures enabling quick completions of transactions.
- With its narrow uni-directional links based on differential signaling, the Intel® QuickPath Interconnect is able to achieve substantially higher signaling rates, thereby delivering the processor interconnect bandwidth necessary to meet the demands of future processor generations.



Interconnect Overview



Block Diagram of Processor with Intel® QuickPath Interconnects

- The physical connectivity of each interconnect link is made up of twenty differential signal pairs plus a differential forwarded clock.
- Each port supports a link pair consisting of two uni-directional links to complete the connection between two components. This supports traffic in both directions simultaneously.
- Various layers are :-
    - The **Physical layer** consists of the actual wires carrying the signals, as well as circuitry and logic to support ancillary features required in the transmission and receipt of the 1s and 0s. The unit of transfer at the Physical layer is 20-bits, which is called a Phit.
    - The next layer up the stack is the **Link layer**, which is responsible for reliable transmission and flow control. The Link layer's unit of transfer is an 80-bit Flit (for Flow control unit).
    - The **Routing layer** provides the framework for directing packets through the fabric.
    - The **Transport layer** is an architecturally defined layer (not implemented in the initial

- Products) providing advanced routing capability for reliable end-to-end transmission.
- The **Protocol layer** is the high-level set of rules for exchanging packets of data between
- Devices. A packet is comprised of an integral number of Flits.

Protocol Agents

The Intel® QuickPath Interconnect coherency protocol consists of two distinct types of agents:

Caching Agent

- A caching agent represents an entity which may initiate transactions into coherent memory, and which may retain copies in its own cache structure.
- The caching agent is defined by the messages it may sink and source according to the behaviour defined in the cache coherence protocol. A caching agent can also provide copies of the coherent memory contents to other caching agents.

Home Agent

- A home agent represents an entity which services coherent transactions, including handshaking as necessary with caching agents.
- A home agent supervises a portion of the coherent memory. Home agent logic is not specifically the memory controller circuits for main memory, but rather the additional Intel® QuickPath Interconnect logic which maintains the coherency for a given address space.
- It is responsible for managing the conflicts that might arise among the different caching agents.
- It provides the appropriate data and ownership responses as required by a given transaction's flow.

Raw Bandwidth

- The raw bandwidth, or maximum theoretical bandwidth, is the rate at which data can be transferred across the connection without any accounting for the packet structure overhead or other effects.
- The Intel® QuickPath Interconnect is a double-pumped data bus, meaning data is captured at the rate of one data transfer per edge of the forwarded clock. So every clock period captures two chunks of data.
- The maximum amount of data sent across a full width Intel® QuickPath Interconnect is 16 bits, or 2 bytes.
- The maximum frequency of the initial Intel® QuickPath Interconnect implementation is 3.2 GHz.

2.6) Intel® Advanced Encryption Standard (AES) Instructions Set

- AES (Advanced Encryption Standard) is an encryption standard adopted by the U.S. government starting in 2001.
- It is widely used across the software ecosystem to protect network traffic, personal data, and corporate IT infrastructure.
- AES is a symmetric block cipher that encrypts/decrypts data through several rounds
- The new 2010 Intel® Core™ processor family (code name Westmere) includes a set of new instructions, **Intel® Advanced Encryption Standard (AES) New Instructions (AES-NI)**.
- The instructions were designed to implement some of the complex and performance intensive steps of the AES algorithm using hardware and thus accelerating the execution of the AES algorithms.
- AES-NI can be used to accelerate the performance of an implementation of AES by 3 to 10x over a completely software implementation.

New Instructions included in AES

- **AESENC**. This instruction performs a single round of encryption. The instruction combines the four steps of the AES algorithm -*ShiftRows, SubBytes, MixColumns & AddRoundKe*y into a single instruction.

- **AESENCLAST**. Instruction for the last round of encryption. Combines the *ShiftRows, SubBytes, & AddRoundKey* steps into one instruction.

- **AESDEC**. Instruction for a single round of decryption. This combines the four steps of AES - *InvShiftRows, InvSubBytes, InvMixColumns, AddRoundKey* into a single instruction

- **AESDECLAST**. Performs last round of decryption. It combines *InvShiftRows, InvSubBytes, AddRoundKey* into one instruction.

- **AESKEYGENASSIST** is used for generating the round keys used for encryption.

- **AESIMC** is used for converting the encryption round keys to a form usable for decryption using the Equivalent Inverse Cipher.

How it Works?

- The AES algorithm works by encrypting a fixed block size of 128 bits of plain text in several rounds to produce the final encrypted cipher text.
- The number of rounds (10, 12, or 14) used depends on the key length (128b, 192b, or 256b).
- Each round performs a sequence of steps on the input state, which is then fed into the following round.
- Each round is encrypted using a subkey that is generated using a key schedule.

Benefits of using AES-NI

A) Performance Improvement

- The performance improvement expected with the use of AES-NI would depend on the applications and how much of the application time is spent in encryption and decryption.
- At the algorithm level, using AES-NI can provide significant speedup of AES
- For non-parallel modes of AES operation such as CBC-encrypt AES-NI can provide a 2-3 fold gain in performance over a completely software approach.
- Performance results for serial and parallel modes of operation are provided for all key sizes, for variable numbers of cores and threads.

B) Improved Security

- Beyond improving performance, the new instructions help address recently discovered side channel attacks on AES.
- AES-NI instructions perform the decryption and encryption completely in hardware without the need for software lookup tables. Therefore using AES-NI can lower the risk of side-channel attacks as well as greatly improve AES performance.

Applications Using AES-NI

- AES is very widely used in applications such as network encryption, disk and file encryption apps.
- File-level and disk encryption applications use AES to protect data stored on a disk. Networking applications use encryption to protect data in flight with protocols encompassing SSL, TLS, IPsec, HTTPS, FTP, SSH, etc.
- There are several ways to take advantage of AES-NI in your applications, whether you are starting from scratch or optimizing existing applications. The following shows several methods for using AES-NI.

  I. Using Standard Libraries
  II. Using C/C++ or assembly

2.7) Intel® Streaming SIMD Extension (SSE)

Introduction

- In computing, **Streaming SIMD Extensions** (**SSE**) is a SIMD instruction set extension to the x86 architecture, designed by Intel and introduced in 1999 in their Pentium III series processors as a reply to AMD's 3DNow!
- SSE contains 70 new instructions, most of which work on single precision floating point data. SIMD instructions can greatly increase performance when exactly the same operations are to be performed on multiple data objects.
- Intel's first IA-32 SIMD effort was the MMX instruction set.
- MMX had two main problems: it re-used existing floating point registers making the CPU unable to work on both floating point and SIMD data at the same time, and it only worked on integers.
- SSE floating point instructions operate on a new independent register set (the XMM registers), and it adds a few integer instructions that work on MMX registers.
- The addition of integer support in SSE2 made MMX largely redundant, though further performance increases can be attained in some situations by using MMX in parallel with SSE operations.
- Building on the already rich Intel® 64 instruction set architecture (ISA), these new instructions will deliver superior performance and energy efficiency to a broad range of 32-bit and 64-bit applications.

Streaming SIMD Extensions 4 (SSE4)

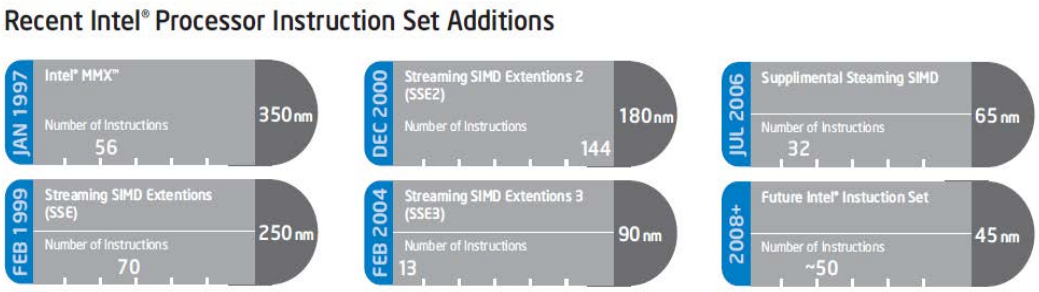
- That will provide building blocks for delivering expanded capabilities, enhanced performance, and greater energy efficiency for most applications.

Application Targeted Accelerators

- That will provide a new foundation for delivering low latency, lower power fixed-function capabilities for targeted applications.
- Applications that will benefit include those involving graphics, video encoding and processing, 3-D imaging, gaming, web servers, and application servers.
- High performance applications that will benefit include data mining; database; complex searching and pattern matching algorithms; audio, video, image, and data compression algorithms; parsing and state machine-based algorithms; and many more.

Enhancement of ISA

- SIMD was a technique employed by Intel to achieve increased parallelism in the P5 microarchitecture through the use of special instructions that operated on multiple pieces of data simultaneously.
- Using Intel MMX technology instruction set, programmers had the ability to execute instructions on multiple data elements loaded into MMX technology registers that would deliver increased performance in media applications such as graphics, gaming, streaming video, and more.
- In the P6 microarchitecture, Intel introduced Streaming SIMD Extensions (SSE). Designed for the Intel® Pentium® III processor, SSE extended MMX technology and allowed SIMD computations to be performed on four packed single-precision FP data elements simultaneously using 128-bit registers (named XMM0-XMM7).
- SSE2 provided the ability to perform more computations in parallel by extending those instructions introduced in MMX technology and SSE, and enabling support of128-bit integer and packed double-precision FP data types. In all, SSE2 added 144 instructions that delivered performance increases across a broad range of applications.
- For instance, SSE2 instructions gave software developers maximum flexibility in implementing algorithms and providing performance enhancements to game and videos.
- The launch of the 90 nm process-based Pentium 4 processor saw the introduction of SSE3. SSE3 includes 13 additional SIMD instructions over SSE2 that are primarily designed to improve thread synchronization and x87-FP math capabilities.
- A further advancement, Supplemental SSE3, is now available in Intel Core microarchitecture. Included in Intel® Xeon® 5100 processors (server and workstation) and the Intel Core 2 Duo processors (notebook and desktop) processors, Supplemental SSE3 adds 32 new opcodes—including align and multiply-add—for yet greater performance.



Overview of SSE4 for Intel Architecture

a. **SSE4 Vectorizing Compiler and Media Accelerators**
   i. SSE4 adds several new compiler vectorization primitives that extend the capabilities of Intel architecture by enabling performance optimized and lower power code generation.
   ii. Applications that will benefit include those involving image processing, graphics, video processing, 2-D/3-D generation, multimedia, gaming, memory-intensive workloads, HPC workloads, and more.

b. **SSE4 Efficient Accelerated String and Text Processing**
   i. SSE4 provides new string and text processing instructions that will enhance the performance of string and text processing operations, resulting in a performance boost for a wide variety of data processing, search, and other text-based applications.
   ii. These new instructions will include advanced packed string comparison instructions that can perform multiple compare and search operations in a single instruction
   iii. Applications that will benefit include those involving databases, text search, virus scanning, string process libraries like ZLIB, Token parsing/recognizing applications like compilers, and state machine oriented applications.

Latest Versions

- SSE2 introduced with the Pentium 4, is a major enhancement to SSE. SSE2 adds new math instructions for double-precision (64-bit) floating point and also extends MMX integer instructions to operate on 128-bit XMM registers.
- SSE3, also called Prescott New Instructions (PNI), is an incremental upgrade to SSE2, adding a handful of DSP-oriented mathematics instructions and some process (thread) management instructions.
- SSSE3 is an incremental upgrade to SSE3, adding 16 new instructions which include permuting the bytes in a word, multiplying 16-bit fixed-point numbers with correct rounding, and within-word accumulate instructions. SSSE3 is often mistaken for SSE4 as this term was used during the development of the Core microarchitecture.
- SSE4 is another major enhancement, adding a dot product instruction, additional integer instructions, a popcnt instruction, and more..
- AVX (Advanced Vector Extensions) is an advanced version of SSE announced by Intel featuring a widened data path from 128 bits to 256 bits and 3-operand instructions. Intel released processors in early 2011 with AVX support.

2.8) Intel® Quick Sync Video

- Intel® Quick Sync Video adds hardware-based video encode acceleration to 2nd generation Intel® Core™ i3, Core™ i5, and Core™ i7 processors,
- Complementing its already integrated, hardware-based decoding engine and eliminating the need for software-based transcoding of video streams or additional hardware.
- With Intel® Quick Sync Video, 2nd generation Intel® Core™ processors integrate a fully accelerated video decode, pre-processing, and encode pipeline for fast video transcoding.

Intel® Quick Sync Video complements the existing Intel® Clear Video HD Technology and its hardware-based video decoding within 2nd generation Intel® Core™ processors

2.9) Intel® Clear Video HD Technology

How it works?

- Intel Clear Video HD Technology integrates a wide range of image-processing technologies, including:

  - Advanced video technologies that remove jitter and create crisper visuals
  - Total color control, along with adaptive contrast and skin-tone enhancements that create vivid, rich colors on the display
  - Intelligent color space mapping that helps ensure that colors are presented the way they were meant to be seen when you are watching video, browsing the Web, or using your PC applications

- **It supports** :

  - Optimized high-definition playback with support for Blu-ray* with dual video hardware decode, letting you watch favorite content in stunning HD
  - Full hardware decode acceleration for AVC, VC1, and MPEG-2 formats, which eliminates stutter and pauses due to software-based video decoding
  - 24-Hz refresh rate support that allows cinematic playback
  - Stunning video quality with sharpness/detail, noise reduction, and Deep Color and xvYCC* that creates incredible visuals
  - High-quality DVD upscaling with Windows Media Player*
  - Advanced de-interlacing and film mode detection to enable crisper imagery
  - Seamless digital display support for HDMI* 1.3 along with dual simultaneous HDMI support that lets users watch content on HDTVs and HD monitors
  - DisplayPort* 1.1 with audio support and up to 2560x1600 resolution that supports Very-high-resolution monitors
  - Multi-channel premium audio that creates amazing audio experiences with native support for bit-streaming Dolby TrueHD* and DTS-HD Master Audio*
  - Up to 8 full range channels at 24-bit/96 kHz

2.10) INTEL CORE MICROARCHITECTURE:

It was implemented in Core 2 duo, Core 2 quad, Core 2 Extreme, Pentium and Celeron desktop processors.

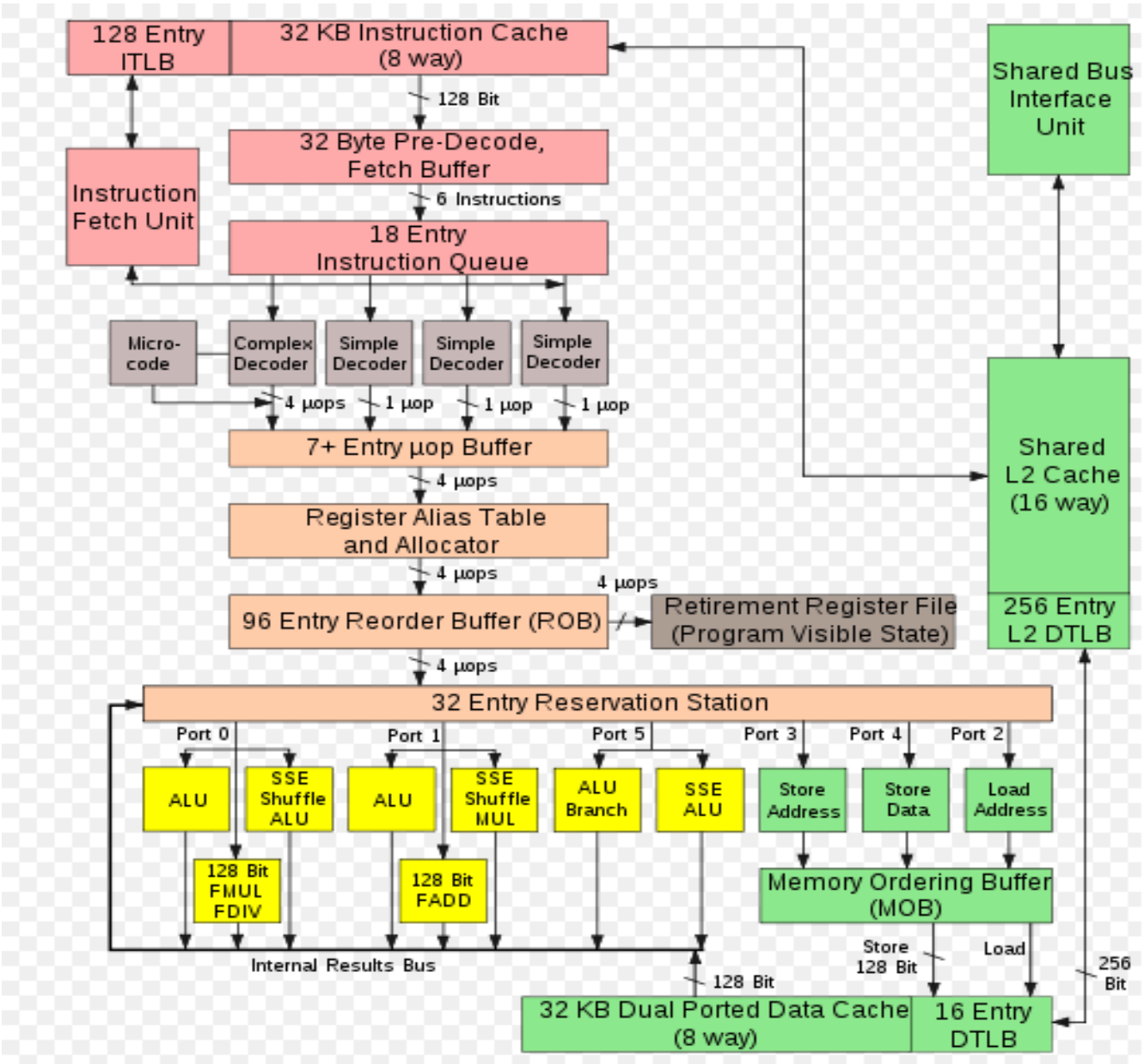


Figure: Core Microarchitecture

FEATURES:

1. The Core microarchitecture returned to lower Clock rate and improved the usage of both available clock cycles and power when compared with the preceding NetBurst Microarchitecture of the Pentium 4-D branded CPUs.
2. The Core microarchitecture provides more efficient decoding stages, execution units, cache, and buses, reducing the power consumption of Core 2-branded CPUs while increasing their processing capacity.
3. Like the last NetBurst CPUs, Core based processors feature multiple cores and hardware virtualization support (marketed as Intel VT-x), as well as Intel 64 and SSE3.
4. However, Core-based processors do not have the Hyper-Threading technology found in Pentium 4 processors. This is because the Core microarchitecture is a descendant of the P6 microarchitecture used by Pentium Pro, Pentium II, Pentium III, and Pentium M.
5. The L1 cache size was enlarged in the Core microarchitecture, from 32KB on Pentium II/III (16 KB L1 Data + 16 KB L1 Instruction) to 64 KB L1 cache/core (32 KB L1 Data + 32 KB L1 Instruction) on Pentium M and Core/Core 2.
6. It also lacks an L3 Cache found in the Gallatin core of the Pentium 4 Extreme Edition, although an L3 Cache is present in high-end versions of Core-based Xeons.

TECHNOLOGIES:

INTEL® WIDE DYNAMIC EXECUTION:

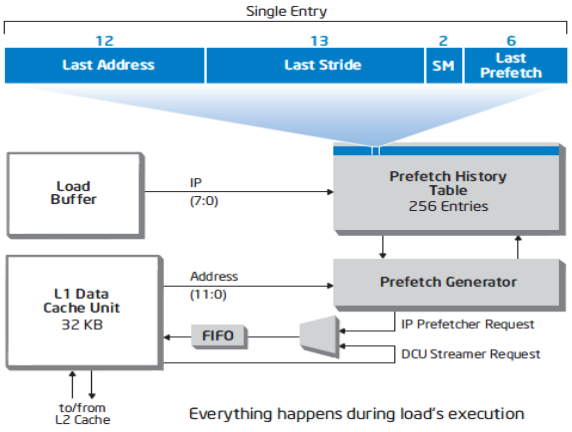
1. In P6 microarchitecture Intel introduced Dynamic Execution (data flow analysis, speculative execution, superscalar). In NetBurst microarchitecture Intel introduced Advanced Dynamic Execution engine to keep executing instruction and an enhanced branch predicting algorithm to reduce misprediction.
2. Core Microarchitecture by having wider execution core can fetch, decode and execute 4 instructions simultaneously.
3. It also incorporate enhanced branch prediction and deeper instruction buffer for greater flexibility.
4. **Macro Fusion**: Common instruction pairs (e.g. Compare followed by jump) are combined in to single internal instruction (micro-op) during decoding. So two instructions become one during executing and overall execution time is reduced.

INTEL® INTELLIGENT POWER CAPABILITY

1. Manages runtime power consumption of all the execution cores.
2. It includes advanced power gating capability that allows for an ultra-fine - grained logic control that turn on logic subsystem only if needed.
3. Many buses and arrays are split so that the data required in some mode of operation can be put in low power state when not needed.

INTEL® SMART MEMORY ACCESS

1. **Memory disambiguation**: Providing each execution core with built in intelligence to speculatively load data for instruction that are about to execute before the previous store instructions are executed.
   - **a.** Normally out of order microprocessors can't reschedule loads before the store instructions because it doesn't know there may be data location dependencies it might be violating.
   - **b.** But many load doesn't depend on the previous store so they can be rescheduled .
   - **c.** In memory disambiguation uses special intelligent algorithms to evaluate whether load can be executed before the preceding store instruction.
   - **d.** If it evaluates valid then load is executed before store to achieve instruction level parallelism and high performance as it doesn't wait for store to execute.
   - **e.** If it evaluates invalid then it is built in with speculative intelligence to find conflict and reload instruction with correct data and execute it.
2. **Prefetchers:** Prefetching the memory content before they are requested so that they can be placed on cache and accessed when needed.
   - **a.** To ensure data is where each execution core needs it Intel uses two Prefetchers per L1 cache and two prefetcher per L2 cache.
   - **b.** This prefetcher detects multiple streaming and strided access patterns simultaneously. This enable to ready data in L1 cache for just in time execution.
   - **c.** L2 cache prefetcher analyzes the access of the core so as to ensure that it contains the data it will need in future.



INTEL® ADVANCED SMART CACHE

1. Details :
    a. L1 & L2 Cache, write back and non-inclusive.
    b. L1 Cache
        i. Instruction Cache – 32 KB, 8 way
        ii. Data Cache – 32 KB, 8 way
    c. L2 Cache (shared) - 6 MB , 24 way
    d. Data TLB (Page size – 4k or Large Pages)
        i. DTLB0 (loads) (256 entries – 32 for large pages)
        ii. DTLB1 (stores & load miss.) (16 entries)
    e. Out of order Loads (buff - 32 entries) & Stores (buff - 20 entries)
2. Cache memory is shared between two core so as to optimize the cache hit as each core can access 100% L2 cache.

INTEL® ADVANCED DIGITAL MEDIA BOOST

1. With the new Core Architecture Intel's processors will be able to execute all of the SSE instructions up to 128-bits in a single cycle.
2. Now the processor could theoretically handle a 128-bit Multiply, Add, Store and Load and still have a single pipeline left open for any other operation. In this diagram Intel showed the macro-fused CMPJCC as the extra instruction being executed so you are essentially getting six x86 instructions done per clock at a maximum.

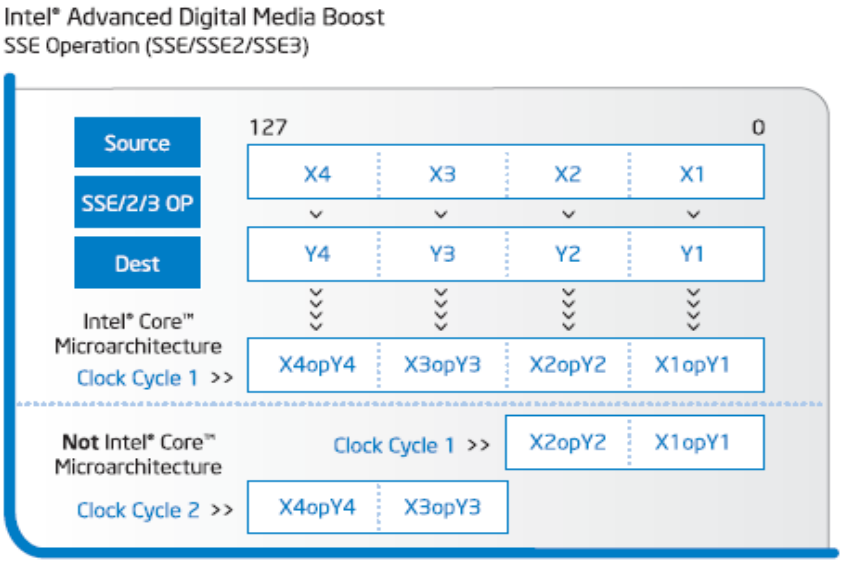
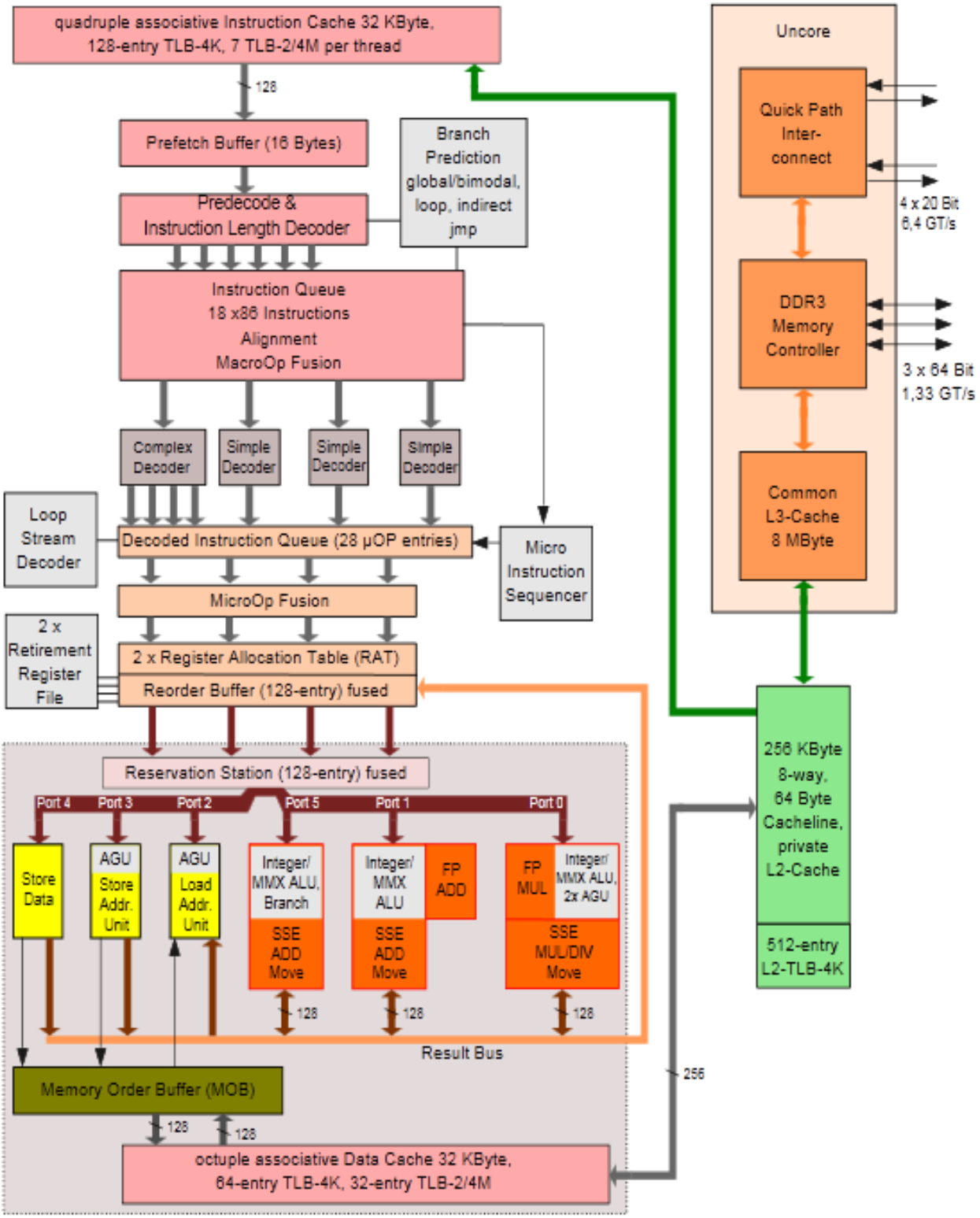


Figure: Example showing advantage of SSE

INTEL NEHALEM MICROARCHITECTURE:

It was implemented in the Core i3, Core i5, Core i7, Pentium and Celeron desktop processor.



FEATURES:

1. Hyper-threading reintroduced.
2. 4–12 MB L3 cache
3. Second-level branch predictor and translation lookaside buffer
4. Native (all processor cores on a single die) quad- and octo-core processors
5. Intel QuickPath Interconnect in high-end models replacing the legacy front side bus
6. 64 KB L1 cache/core (32 KB L1 Data + 32 KB L1 Instruction) and 256 KB L2 cache/core.
7. Integration of PCI Express and DMI into the processor in mid-range models, replacing the Northbridge
8. Integrated memory controller supporting two or three memory channels of DDR3 SDRAM or four FB-DIMM2 channels
9. 2nd generation Intel Virtualization Technology which introduced Extended Page Table support, virtual processor identifiers (VPIDs), and non-maskable interrupt-window exiting

PERFORMANCE:

1. It has been reported that Nehalem has a focus on performance, thus the increased core size.
2. Compared to Penryn, Nehalem has 10-25% more single-threaded performance / 20-100% more multithreaded performance at the same power level 30% lower power usage for the same performance
3. Nehalem provides a 15–20% clock-for-clock increase in performance per core(average)
4. The Nehalem architecture reduces atomic operation latency by 50% in an attempt to eliminate atomic

ARCHITECTURE:

Core architecture was already very efficient performing so there were some improvements made in that all the rest is same.

a) Advanced Processor Core :

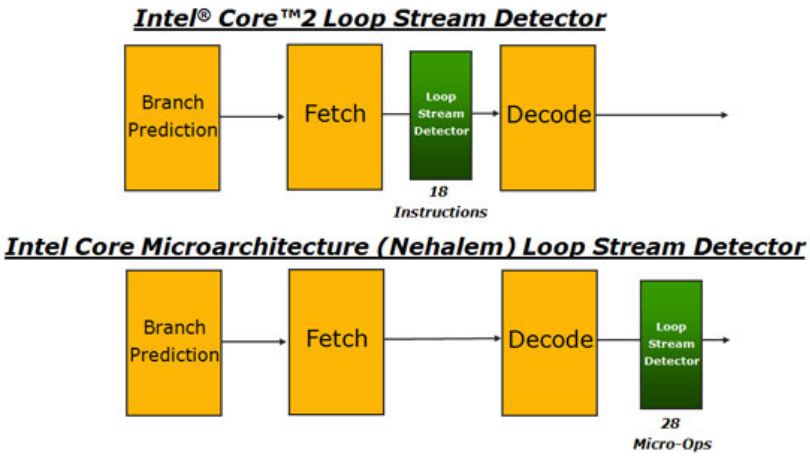
- **Improved Macrofusion:**

  i. Processors with Core microarchitecture had four decoders .Macrofusion technology allowed Core 2 processors to process certain pairs of instructions as a single command - for example, comparison followed by conditional branching.
  ii. Nehalem has the same number of the same decoders. However, Macrofusion technology did change significantly.
    1. There are more pairs of x86 instructions decoded “at one fling” within this technology.
    2. Macrofusion technology in Nehalem processors works in 64-bit mode, while in Core 2 processors it could only be activated when the CPU worked with 32-bit code.

  CPUs with new microarchitecture will be able to decode five instructions per clock instead of four in a larger number of cases than their predecessors.

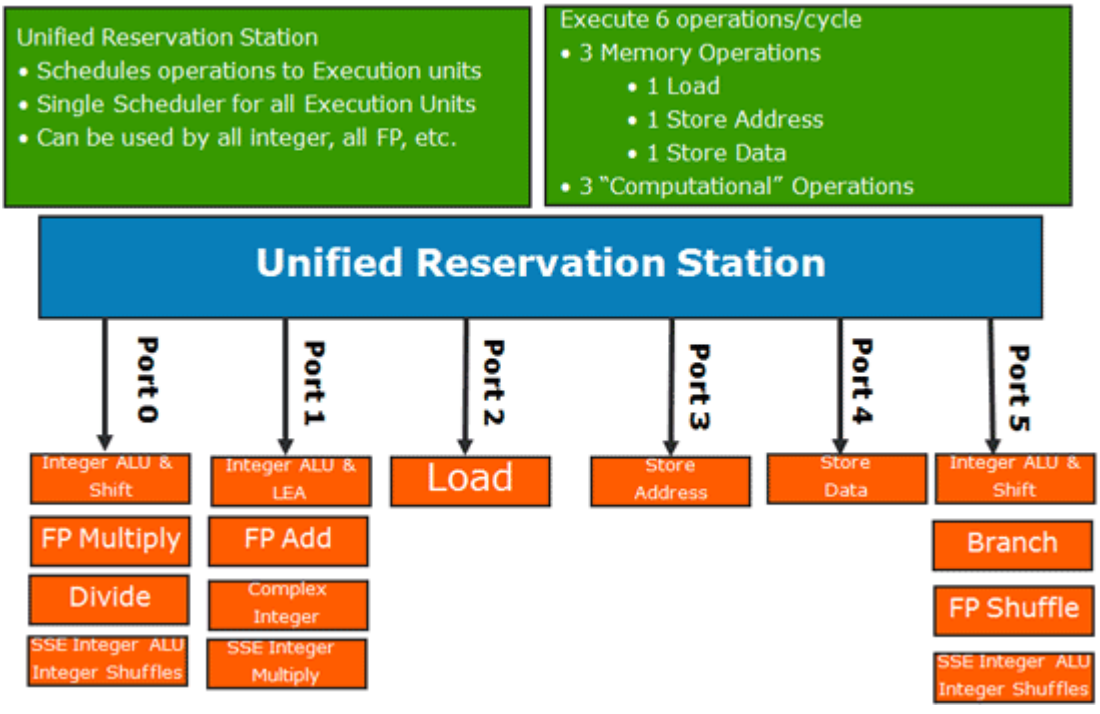
- **Loop Stream Detector block** :

  i. This block first appeared in CPUs with Core microarchitecture and was designed to speed up loops processing. Loop Stream Detector detected small loops in the program code and saved them in a special buffer. As a result, the CPU didn’t have to fetch them from the cache over and over again and predict branching within these loops.
  ii. Nehalem processors have an even more efficient Loop Stream Detector block, which has been moved past the instructions decoding stage. In other words, Loop Stream Detector now saves decoded loops, which makes it a little similar to Trace Cache of Pentium 4 processors.
  iii. However, Loop Stream Detector of Nehalem CPUs is a specific cache.
    - It is very small, only 28 micro-ops.
    - It saves only loops.



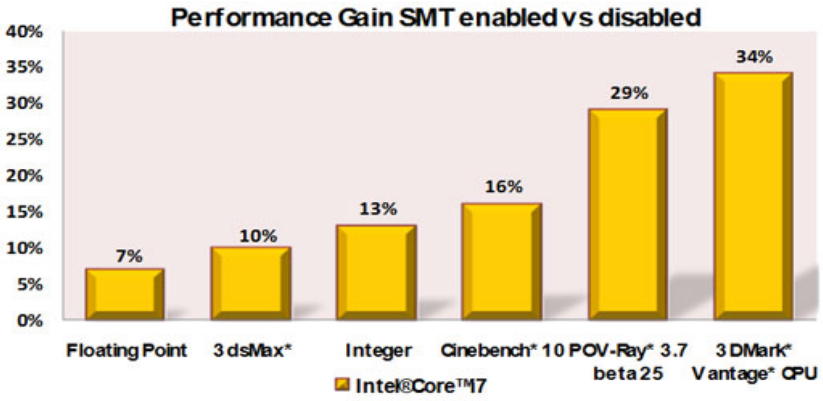
- **Second Level branch predictor** :
  - i. They simply added one more second-level predictor to the already existing branch prediction unit. It is slower than the first one, but features a larger buffer for storing the branching statistics and hence boasts more analysis depth.
  - ii. They also improved the efficiency of the branch prediction unit by changing ***Return Stack Buffer*** unit. This unit is responsible for correct prediction of functions return addresses. However, previous generation processors could predict function return addresses incorrectly, for example when recursive algorithms were working and the corresponding buffer got overfilled. The new Return Stack Buffer implemented in Nehalem processors didn't have this problem anymore.

- **Execution unit** :
  - i. The execution units were almost same as Core micro architecture.
  - ii. Like Core 2, CPUs on Nehalem microarchitecture can send up to 6 micro-operations at a time for processing.
  - iii. However, the developers have increased the size of the buffers on commands execution stage.
    - As a result, Nehalem processors can hold up to 128 micro-ops waiting to be executed in the **Reorder Buffer**, which is 33% more than Core 2 can.
    - As a result, **Reservation Station** sending micro-operations directly to execution units has been increased from 32 to 36 instructions. They have also made the data buffers larger.



b) **Simultaneous Multi-Threading** :

- Processor can simultaneously process up to two computational threads that require resource sharing.
- Pentium 4 processors where the exact same technology was presented as Hyper-Threading, received up to 10% average performance boost from enabling it.
- Processors with Nehalem microarchitecture should benefit even more from SMT.
    i. It has memory subsystem with much higher bandwidth that can much better supply two computational processes with data.
    ii. Nehalem boasts “wider” microarchitecture that allows processing more instructions simultaneously.



- When SMT is enabled, all other resources are shared dynamically between processor threads (for example, Reservation Station or cache-memory) or shared 50-50 (for example, Reorder Buffer).
- Enabling SMT in Nehalem makes the operating system see each physical processor core as a pair of logical cores.
- Compare Nehalem against Penryn in a few simple benchmarks from SiSoftware Sandra 2009 suite.

Sandra 2009	Nehalem 3.2 GHz SMT Enable	Nehalem 3.2 GHz SMT Disable	Penryn 3.2 GHz
Dhrystone ALU, GIPS	81.44	71.03	54.98
Whetstone FPU, GFLOPS	72.09	45.44	47.15
Multimedia Int x8, Mpixel/s	143.07	113.98	118.6
Multimedia Float x4, Mpixel/s	115.33	73.89	65.94
Multimedia Double x2, Mpixel/s	59.21	36.72	33.67

- In above example enabling SMT improves Nehalem results by 15-60%. However, is we compare the results of Nehalem and Penryn processors without SMT, then the new processor will not always be better than its predecessor. Everything depends on the type of workload, which indicates that there have been no revolutionary or universal changes made to the new core.

c) **TLB and Cache memory** :

- TLB :

    - They significantly increased the size of the TLB (Translation-Lookaside Buffer). As you know, TLB is a high-speed buffer used to map over the physical and virtual page addresses. By making the TLB bigger they increase the number of memory pages that can be used without additional costly modifications employing address translation tables stored in regular memory.
    - Moreover, TLB of Nehalem processors became dual-level. In fact, Intel simply added another L2 buffer to the TLB inherited from Core 2 processors. The new L2 TLB is not only large and can save up to 512 entries, but also boasts relatively low latency. Also, the new L2 TLB is unified and can translate page addresses of any size.
    - It is evident that TLB modifications were intended primarily for server applications that require a lot of memory. However, the increased number of TLB entries may also have a positive effect on the memory subsystem performance in desktop tasks, too.
    - Especially since both TLB levels are dynamically shared between the virtual cores when SMT technology is enabled, so the opportunity to save additional entries in this buffer will not go to waste.

- Another innovation that should increase the memory subsystem performance in CPUs on Nehalem microarchitecture is significant acceleration of instructions dealing with the data that haven't been aligned along cache-memory lines.
    - They have made first shy attempts to implement it back in Penryn processors, but only in Nehalem CPUs they managed to succeed.
    - Now SSE instructions using 16-byte data successions as operands work equally fast independent of the instruction type: for aligned or unaligned data.
    - Since most compilers translate the code with unaligned instructions, this innovation should definitely improve the performance of applications working with media-content.

- Cache:
    - However, faster processing of unaligned data and adding L2 TLB are trifles compared with the dramatic modification of the cache-memory subsystem in the new Nehalem processors.
    - From the old dual-level cache-memory structure with a shared L2 cache for each pair of cores they only borrowed a 64KB L1 cache split in two equal parts for storing data and instructions.
    - L1 cache in Nehalem processors remained the same, its latency got 1 clock cycle higher than that of the L1 cache in Core 2. It resulted from more aggressive power-saving modes introduced in the new processors that according to Intel have little effect on the overall performance.
    - Although shared L2 cache proved to be highly efficient in CPUs on Core microarchitecture, it appeared pretty difficult to implement in processors with more cores. Therefore, Nehalem microarchitecture allowing processors with up to 8 cores, doesn't have a shared L2 cache any more. Each core gets its own L2 cache of relatively small size: 256KB. However, due to its limited size, the cache boasts lower latency than L2 cache of Core 2 processors. It partially makes up for the higher latency of L1 cache in Nehalem.
    - Nehalem also acquired L3 cache, which connects all cores and is shared. As a result, L2 cache turns into a buffer when processor cores send their requests to pretty big shared cache-memory.

    - Cache Organization :
        - First, L3 cache of the upcoming Intel processors works at higher frequency that will be set at 2.66GHz for the first representatives of this family and may increase later on.
        - The cache-memory remained inclusive, i.e. the data stored in L1 and L2 caches is duplicated in L3 cache. And there is a very good reason for that.
            - Inclusive shared cache speeds up the memory subsystem in multi-core processors due to excessive duplication of L1 and L2 caches of all their cores.
        - If the data requested by one of the cores is not there, it doesn't make sense to look for them in the individual caches of other cores.
        - Since each line in L3 cache has additional flags indicating where this data comes from, the reverse modification of the cache line is also performed fairly simply. If a core modifies the data in L3 cache and these data initially belong to different core/cores, the L1/L2 caches of these cores get updated. This allows eliminating excessive inter-core traffic ensuring coherency of inclusive cache-memory.
        - The results of Nehalem cache-memory latency tests show that this solution is extremely efficient:

Sandra 2009	Nehalem 3.2GHz	Penryn 3.2GHz
L1 Cache Latency	4 cycles	3 cycles
L2 Cache Latency	11 cycles	18 cycles
L3 Cache Latency	52 cycles	-

        - L2 cache of Nehalem processor does in fact have extremely low latency. L3 cache also shows very good access time despite its relatively large size.
        - Despite a dramatic modification of the caching system, Intel didn't change the prefetch algorithms: Nehalem has borrowed them as is from Core 2.

d) **New SSE 4.2 instructions** :

- Intel continued increasing the number of supported SIMD instructions in their new Nehalem microarchitecture.
- They added a set of seven new instructions called SSE4.2 (they used SSE4.1 name for SIMD instructions of Penryn CPUs). Intel specifically stressed that the new SSE4.2 instructions are designed not that much for the processing of streaming media content, but for slightly different things. Therefore, new Nehalem instructions are also called ATA (Applications Targeted Accelerators).
- ATA concept implies that contemporary technological processes allow employing some processor transistors not only in universal functional units but also in some specific tasks increasing the performance in particular applications.
- According to this concept, SSE4.2 has five instructions accelerating XML-files parsing. These instructions also speed up lines and texts processing.
- Another two new instructions from SSE4.2 set are intended for completely different applications.
    i. CRC32 accumulates checksum
    ii. POPCNT, counts the number of set bits in the source.

These instructions can also be used for a wide range of crunching and networking tasks.

e) **Integrated memory controller** :

- The main feature of memory controller is its flexibility. Keeping in mind the modular design of the entire upcoming processor family, that may include solutions differing dramatically in features and market positioning, Intel foresaw the opportunity not just to enable or disable buffered modules, but also to vary the memory speed and the number of channels.
- The first processors with Nehalem microarchitecture will be quad-core models and they will have a triple-channel memory controller supporting DDR3 SDRAM. This way, desktop systems built on new CPUs will boast unprecedented memory subsystem bandwidth. With three DDR3-1067 SDRAM modules it will reach 25.6GB/s.
- However, the main advantage of transferring the DRAM controller into the CPU is not the bandwidth increase, but lowering of the memory subsystem latency. Although Intel designed their new processors to work with DDR3 SDRAM that has relatively high latency, it will still be lower that the latency of Core 2 based systems equipped with DDR2 SDRAM.
- Memory subsystem tests performed on a Nehalem based platform in Everest 4.60:

Everest 4.60	Nehalem 3.2GHz			Penryn 3.2GHz	
	3 channel DDR3-1067 7-7-7-20	2 channel DDR3-1067 7-7-7-20	1 channel DDR3-1067 7-7-7-20	2 channel DDR3-1067 7-7-7-20	2 channel DDR3-1600 7-7-7-20
Read, MB/s	15057	14030	8199	8125	10101
Write, MB/s	14727	12038	8200	8464	8495
Copy, MB/s	15522	13314	8560	7207	8175
Latency, ns	39.2	34.9	33.5	64.2	51.7

- Even in single-channel mode, Nehalem memory controller performs better than chipset memory controller in LGA775 platforms. It is an absolutely logical result, because there are no intermediate devices between the CPU and the memory in the new generation processors. However, before that the chipset North Bridge was responsible for work with the memory subsystem and since it had to synchronize the memory bus and the FSB, it did affect the memory subsystem latency.
- Another indirect advantage of the built-in memory controller is its complete independence of the chipset and the mainboard. As a result, Nehalem will work with the memory subsystem equally fast in platforms from different developers.

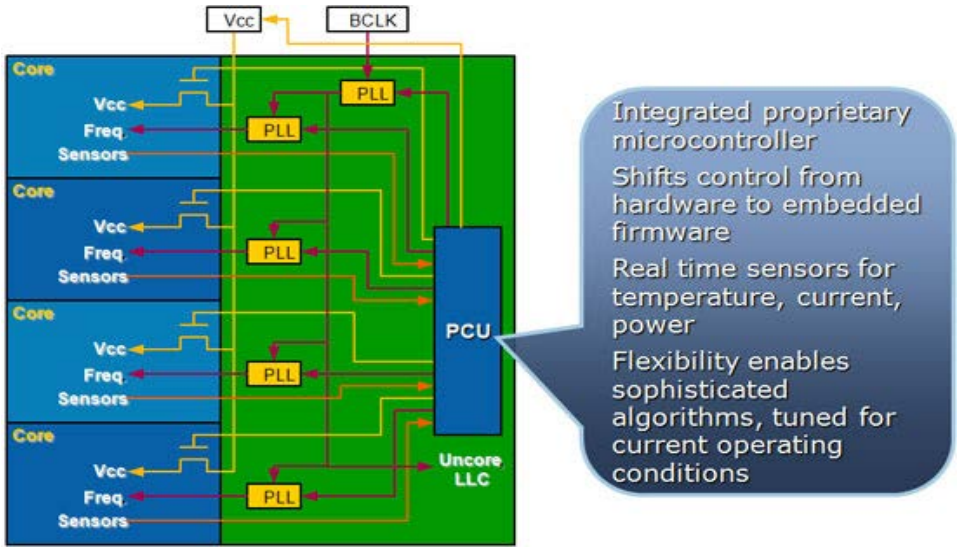
f) Quick Path Interconnect Bus :

- You may believe that by moving the memory controller into the CPU, they should have taken a lot of load off the processor bus that is this case doesn't have to transfer data between the CPU and the memory any more. It is partially true, but only for single-processor systems. Nehalem microarchitecture is universal; it should be used for desktop and mobile as well as server solutions. That is why Intel designed a new processor bus that could suit for multi-processor systems and provide sufficient bandwidth and scalability.
- The traditional FSB bus cannot be used in this case. Multi-processor systems on processors with integrated memory controllers should use NUMA memory model (Non-Uniform Memory Access) and hence require direct high-speed connection between the CPUs.
- To accomplish this task they built special serial interface called CSI (Common System Interface) with point-to-point topology that was later renamed to QPI (Quick Path Interconnect).
- On the technical side, QPI consists of two 20-bit links transferring data forward and back. 16 bit are assigned for data and the remaining 4 bits serve some auxiliary purpose: they are used by the protocol and error correction. This bus performs maximum 6.4 mln transfers per second (GT/s) and has 12.8GB/s bandwidth in each direction, or 25.6GB/s total bandwidth.
- The current bandwidth of the new QPI bus allows us to call it the fastest processor bus out there. The old Quad Pumped Bus can only reach 12.8GB/s total bandwidth at 1600MHz frequency. Hyper Transport 3.0 bus similar to QPI and used in contemporary AMD processors can boast only 24GB/s peak bandwidth.
- E.g.

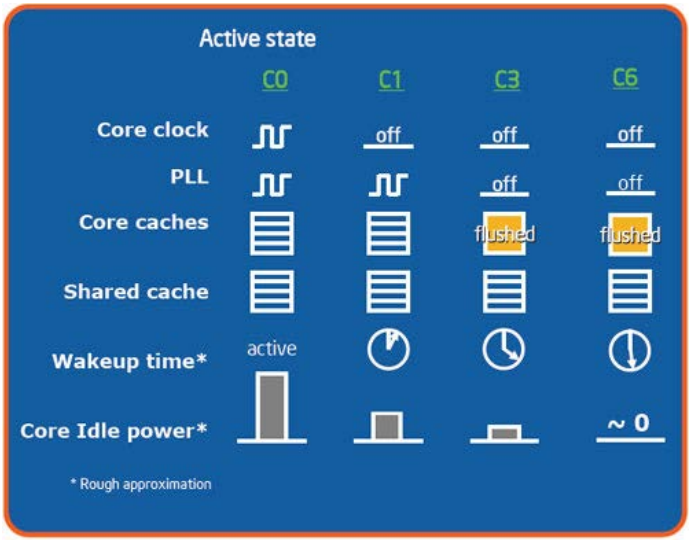
Core i7 Extreme	990X	3.46 GHz	1× QPI 6.4 GT/s
	980X	3.33 GHz	
Core i7	980		1× QPI 4.8 GT/s
	970	3.20 GHz	

g) Power management and Turbo Mode :

- Multi-core processors on Core microarchitecture are very power-inefficient in the sense that there is a single algorithm for their power management needs that doesn't take into account the individual cores. Therefore, it is a pretty frequent situation when one of the cores in contemporary quad-core CPUs that is loaded heavily prevents other cores from going into power-saving mode even though they are hardly involved.
- Nehalem microarchitecture has one more important processor unit called PCU (Power Control Unit). It is actually just another programmable micro-controller built into the CPU that should manage power consumption intelligently.



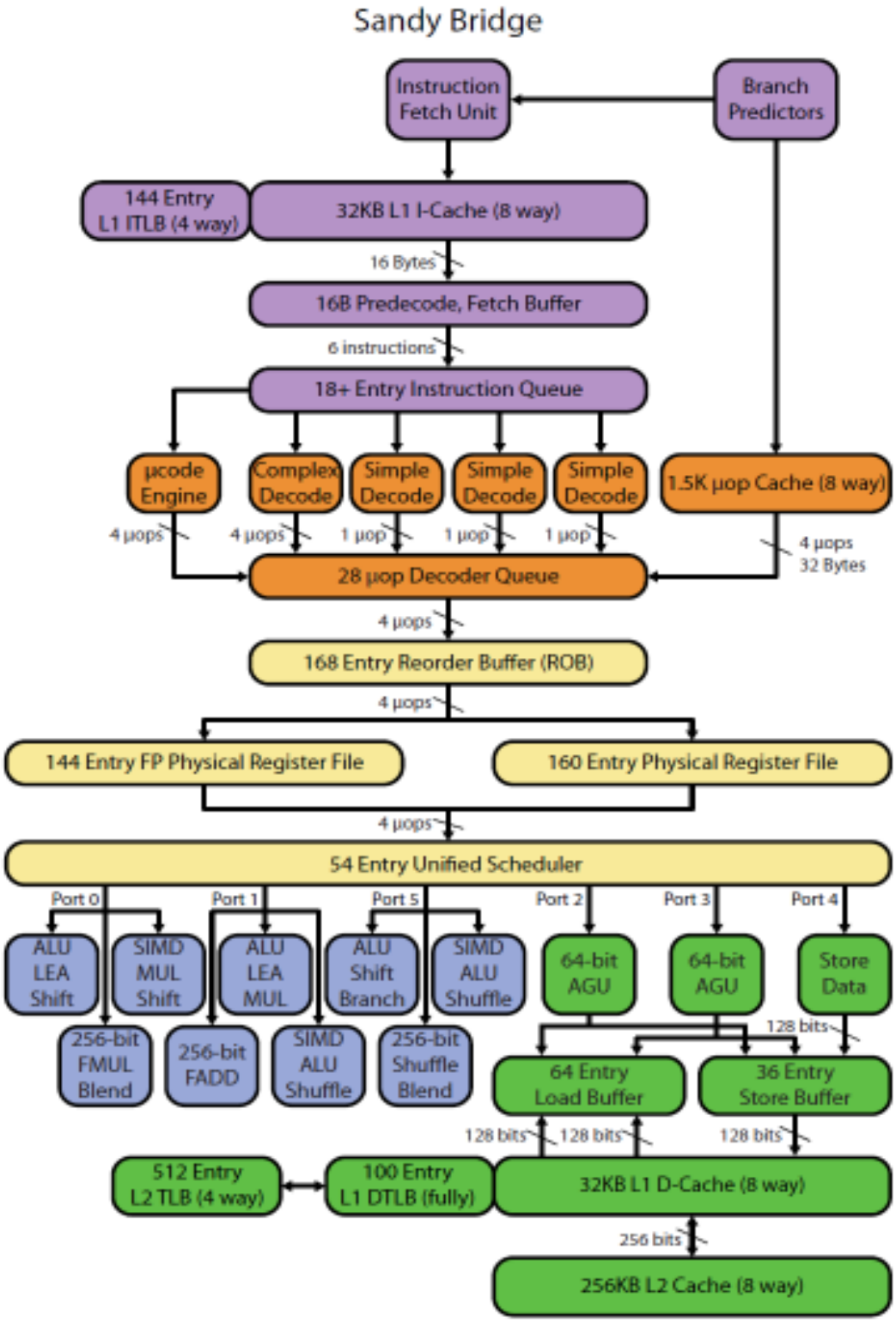
- PCU's main task is to adjust the frequency and voltage of individual cores and it has everything it takes for that.
- It receives the sensor readings of temperatures, voltage and current for all cores. PCU analyzes these data and switches qualifying cores to power-saving mode by adjusting their frequency and voltage. Namely, PCU may disable inactive cores and put them in
  Deep sleep state where their power consumption will be close to 0.



- Intel created special semiconductor material that allowed disconnecting the cores from the power bus independently. The main advantage of this technology is that power management of individual cores is performed inside the CPU and doesn't require enhancing the processor voltage regulator circuitry on mainboards in any way.
- As for the processor units identical for all cores, such as memory controllers and QPI interface, they go into power-saving mode when all processor cores sleep.
- An intelligent controller that can manage processor cores independently allowed Intel to implement one more interesting technology called Turbo Boost Technology. It introduces a Turbo-mode, when individual cores can work at frequencies exceeding the nominal, i.e. be overclocked. According to Turbo Boost Technology main principle, the overall processor power consumption and heat dissipation lowers when some cores go into power-saving mode, which allows increasing the frequencies of other cores without risking of getting past the TDP limits.
- Turbo-mode doesn't necessarily get enabled when one or more cores go into power-saving mode. It is simply one of the possible scenarios. Since the PCU can get all the information on the current processor cores status, Turbo-mode can also be enabled when all cores are active but the workload is relatively small.
- Turbo Boost Technology is absolutely transparent for the operating system and it is its great advantage. It is implemented only in hardware and doesn't require any software applications or utilities to be running.
- In core i7 extreme processors Intel Enhanced SpeedStep technology kicks in when there is no workload: the processor frequency drops to 1.6GHz. Launching one thread activated one core, so the CPU can increase its multiplier from 24x to 26x overclocking itself to 3.46GHz. Two threads increase the processor load so much, that PCU only dares raise the clock speed to 3.33GHz. The frequency remains at this point until we have 5 simultaneous threads working. And only sixth thread increases the CPU utilization to 75% lowering its frequency back to the nominal 3.2GHz. In other words, Turbo Boost Technology is not an ephemeral concept: its effect is real.

INTEL SANDY BRIDGE MICROARCHITECTURE:

It is created using 32 nanometer manufacturing process based on planar double-gate transistors. This Microarchitecture will have everything integrated into a single chip: processor cores, graphics core, memory controller and PCI Express bus controller.



FEATURES:

- 32 kB data + 32 kB instruction L1 cache (3 clocks) and 256 kB L2 cache (8 clocks) per core
- Shared L3 cache includes the processor graphics (LGA 1155) 64-byte cache line size
- Two load/store operations per CPU cycle for each memory channel
- Decoded micro-operation cache and enlarged, optimized branch predictor
- Improved performance for transcendental mathematics, AES encryption (AES instruction set), and SHA-1 hashing
- 256-bit/cycle ring bus interconnect between cores, graphics, cache and System Agent Domain
- Advanced Vector Extensions (AVX) 256-bit instruction set with wider vectors, new extensible syntax and rich functionality
- Intel Quick Sync Video, hardware support for video encoding and decoding
- Up to 8 physical cores or 16 logical cores through Hyper-threading

PERFORMANCE:

- If we compare Sandy Bridge and Lynnfield processors with the same number of computational cores that work at the same clock frequency, the new microarchitecture delivers 5-10% higher actual performance.
- At the same time, the power consumption of Sandy Bridge processors appears about 25% lower, i.e. the new CPUs have progressed substantially in terms of performance-per-watt ratio.
- The nominal clock speeds of Sandy Bridge processors are about 10% higher than those of comparable Lynnfield ones, we can conclude that the new platform as a whole will be at least 25% faster than the previous-generation LGA1156 platform.

ARCHITECTURE:

1. Pipelining:

- **Decoded micro-op cache:**
  - The actual decoder remained the same as in Nehalem – it processes 4 instructions per clock cycle and supports Micro-Fusion and Macro Fusion technologies that make the output instructions thread more even in terms of execution complexity.
  - However, the processor instructions decoded into micro-operations are not just transferred to the next processing stage, but also cached. In other words, in addition to the regular 32 KB L1 cache for instructions Sandy Bridge also has an additional "L0" cache for storing the decoding results. This cache is the first flashback from NetBurst microarchitecture, its general operation principles make it similar to the Execution Trace Cache.
  - The decoded micro-ops cache is about 6 KB big and can store up to 1500 micro-ops, which makes it of great help to the decoder.
  - If the decoder discovers instructions that have been translated earlier and are now stored in the cache, it replaces them with internal micro-operations without performing any new decoding. This decoded micro-ops cache helps to take a big load off the decoder, which is a pretty energy-hungry part of the CPU.
  - According to Intel, this additional cache comes in handy in about 80% of cases, which makes all suspicions about its inefficiency absolutely unjustified. Besides, when the decoder in Sandy Bridge is idle, it is disabled thus helping lower the CPU power consumption substantially.
- **Enhanced branch prediction unit :**
  - Intel modified all the Sandy Bridge buffers used to store branch addresses and prediction history in order to increase the data density in them. As a result, Intel is able to store longer branching history without increasing the size of the data structures used by the branch prediction unit.
  - According to preliminary estimates, the branch prediction correctness in Sandy Bridge improved by more than 5% compared with the predecessor.
- **Out-of-Order cluster :**

- Intel brought back the physical register file into their new processors .Intel retired this file in their Core and Nehalem processors in favor of a centralized Retirement Register File. Before, when they rearranged micro-ops, they used to store full copies of registers for each operation in the buffer.
- Now Intel use links to register values stored in a physical register file. This approach allows not only to eliminate excessive data transfers, but also to prevent multiple duplication of the register contents thus saving space in the register file.

	Nehalem	Sandy Bridge
Load Buffers	48	64
Store Buffers	32	36
RS - Scheduler Entries	36	54
PRF integer	N/A	160
PRF float-point	N/A	144
ROB Entries	128	168

- As a result, the out-of order cluster in Sandy Bridge processors can keep up to 168 micro-ops “in sight” at the same time, while Nehalem processors could store only 128 micro-ops in their ROB (reorder buffer). Besides, some energy is also being saved.
- However, replacing the actual values with the links to them also has its negative side: the execution pipeline gets new stages required for dereferencing the pointers.
- These processors support new AVX instructions operating 256-bit registers, so transferring their values forth and back numerous times would inevitably create additional overhead expenses.
- AVX instructions are none other than further development of SSE, which increases the size of the SIMD vector registers to 256 bit. Moreover, the new instruction set allows non-destructive execution, i.e. when the original data in the registers is not lost. Their implementation will allow simplifying many algorithms and using fewer instructions to complete the tasks.

➢ **Execution units：**

- Intel redesigned EU specifically to ensure that 256 bit instructions can be executed effectively.
- Since each of the three execution ports in Sandy Bridge processors (just like in Nehalem ones) has units for simultaneous work with three types of data – 64 bit, 128 bit integer and 128 bit real – it makes perfect sense to join SIMD units into pairs within the same port. And most importantly, this resources rearrangement doesn’t affect the bandwidth of the processor execution unit at all.

➢ **Memory cluster：**

- In order to increase above three port efficiency Intel unified two of these ports that used to serve for storing addresses and loading data. Now they have become equal and can either load addresses and data or unload addresses.
- The third port remained unchanged and is designed for storing data. Since each port can let through up to 16 bytes per clock, the total throughput of the L1 data cache in the new microarchitecture increased by 50%. As a result, CPUs with Sandy Bridge microarchitecture can load up to 32 bytes of data and store 16 bytes of data per clock cycle.

2. Full Integration and Ring bus technology:

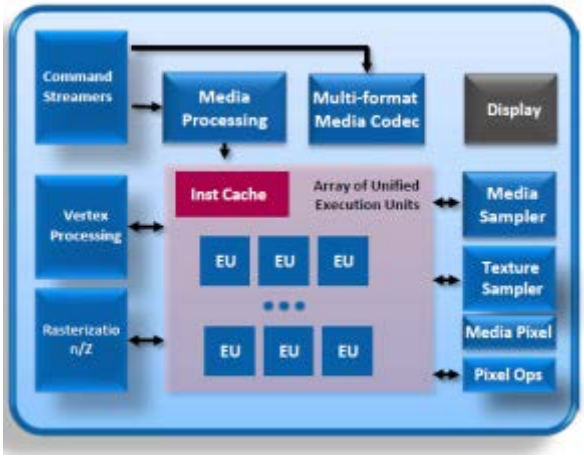
- With the introduction of Nehalem microarchitecture Intel integrated memory controller, PCI Express bus controller, graphics core. The CPUs also received an L3 cache. In other words, the CPU has turned not into just a local computing center, but into a bundle of a numerous different complex units.
- However, the more different units there are inside a CPU, the more difficult it is to connect them all electrically. And the most difficult task in this case would be the connection between shared L3 cache and the processor cores.
- Intel gave a lot of serious thought to convenient connection between all functional units inside the processor. The formerly used common crossbar interconnect worked fine in dual-, quad- and six-core Nehalem processors, but it doesn’t fit for modular CPUs with a large number of different cores inside.
- In fact, they have already taken it into account in eight-core server Nehalem-EX processors where they used an absolutely new technology to connect computational cores with the L3 cache. This technology is called Ring Bus and it has successfully migrated to the new Sandy Bridge microarchitecture.

- All the computational cores, cache, graphics core and North Bridge elements inside the new processors are connected via special ring bus supporting a QPI-like protocol, which allowed to significantly reducing the number of inter-processor connections needed for signal routing.
- They divided the L3 cache of Sandy Bridge processors into equal banks, 2 MB each, in order to ensure communication between the processor functional units with the L3 cache via the ring bus. The original design implies that the number of these banks equals the number of processor cores. However, for marketing reasons they can disconnect some banks from the bus without any damage to the cache integrity and thus reduce the cache-memory size.
- Each of the cache-memory banks is managed by its own arbiter but at the same time all of them work closely together: the data in them is never duplicated.
- The use of banks doesn't split the L3 cache, but rather increases its bandwidth, which in its turn scales according to the growing number of cores, and banks, respectively. For example, since the "ring" is 32 bytes wide, the peak L3 cache bandwidth inside a quad-core CPU working at 3.4 GHz frequency is 435.2 GB/s.
- Advantage :
    - Scalability to the number of processor cores
    - The latency of the L3 cache has also gone down, since data transfers along the "ring" take the shortest route. Now the L3 cache latency is 26-31 clock cycles, while the L3 cache in Nehalem processors offered 35-40 clocks latency. However, in this case you should keep in mind that all cache memory in Sandy Bridge works at the processor frequency, i.e. this is another reason why it has become faster.
    - Also include the graphics core integrated into the processor to the general data transfer routes. It means that the graphics core in Sandy Bridge doesn't work directly with the memory, but like the processor cores do – via L3 cache. This way it works faster and also eliminates the negative effect on the overall system performance caused by the graphics core trying to take part of the memory bus from the processor cores for its own needs.

3. GPU:

➢ GPU architecture modification:

- Overall, graphics core architecture hasn't changed dramatically: it is still based on the same 12 execution (shader) processors.
    - Among the innovative changes are Shader Model 4.1 and DirectX 10.1 support.



- Next Generation EU:

    - Larger register file for increased parallelism and efficient complex shader execution.
    - 2nd Generation parallel branch for efficient parallelization in the face of deeply nested conditions.
    - New transcendental math capability for 4x-20x more throughput
    - New instructions to reach 1 to 1 with API ISA (CISC) and higher throughput at same clock rate.

- Since graphics core has now moved into a 32 nm semiconductor die, it is now possible to easily increase its clock speed, which may go as high as 1.35 GHz. As a result, the performance of Sandy Bridge graphics core in real applications will be comparable with that of entry-level discrete graphics accelerators

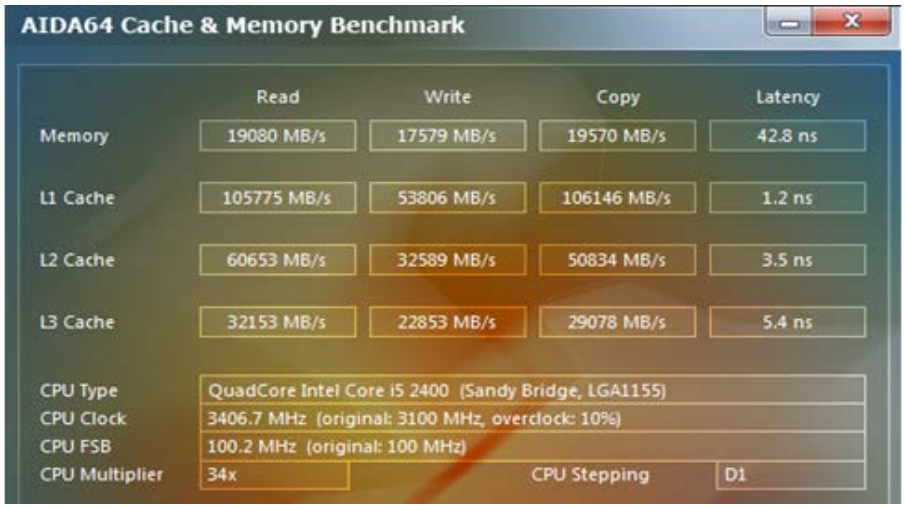
➢ Media Architecture :

- Programmable and Flexible architecture
  - EUs optimized for higher workloads
  - Parallel compute synergy with 3d processing
- Performance Optimized :
  - Native support for mainstream codecs
  - Parallel engine deliver high throughput video rendering
- Dedicated Hardware Accelerators :
  - More compute for HD workloads
  - High quality enhancement and filters
- Special new units intended for video decoding and encoding in popular formats, such as MPEG2, VC1 and AVC.
- Hardware video decoding operation used to be the responsibility of shader processors, while now there is a special functional unit in charge of it. This rearrangement of responsibilities was required for proper support of 3D video: the new graphics core can easily cope with hardware decoding of stereo 3D Blu-ray or MVC streams.
- Another interesting addition is the hardware codec that can encode video into AVC format. In practical terms it means that Sandy Bridge has all the resources necessary for fast video transcoding without utilizing any of the traditional CPU resources for that purpose.
- When multimedia units inside Sandy Bridge graphics core are utilized for video transcoding, shader processors remain free, so they can give a hand in additional video processing or applying special effects.

4. New System Agent:

- System Agent that contains all external interface controllers: PCI Express, DMI, memory and display interfaces.
- System Agent is very similar to what we know as Uncore in Nehalem processors. However, System Agent in Sandy Bridge is nevertheless not fully identical to Uncore.
  - It doesn't contain the L3 cache, which is an individual functional unit working at the processor frequency in the new microarchitecture.
  - It also uses the ring bus to exchange data with the processor and graphics cores as well as with the L3 cache.
- Smart integration with ring provides
  - Core/GPU/Media with high BW ,Low latency to DRAM/IO
  - Handles IO to cache coherency

- Sandy Bridge processors support multipliers that also allow clocking the memory at 1600, 1866 and 2133 MHz.
- You can get an idea of how fast the memory controller in Sandy Bridge processors actually is from the following results of Aida64 benchmark:



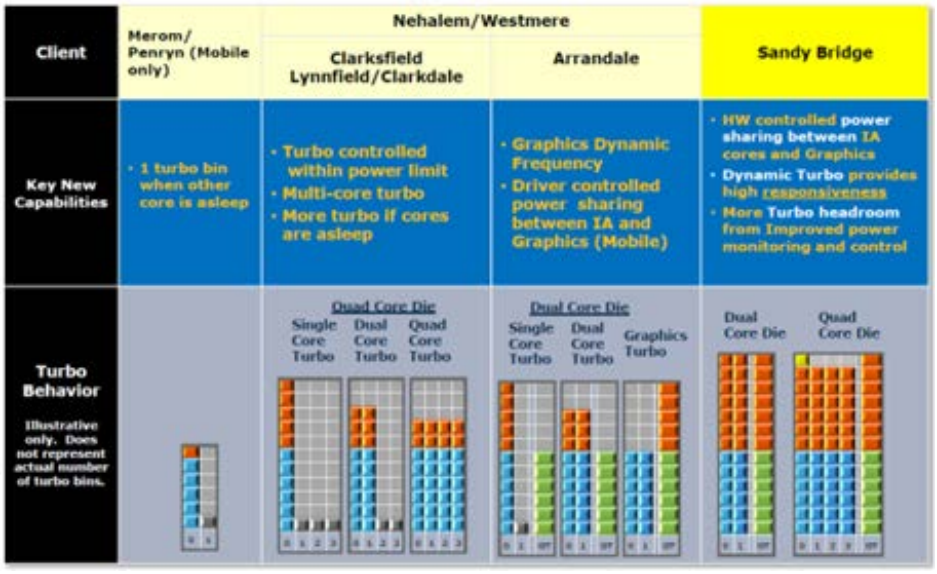
- The PCIE bus controller in Sandy Bridge is similar to the same controller in LGA1156 processors. It supports 16 PCI Express 2.0 lanes that can either form a single PCIE x16 bus, or two PCIE 8x busses. Therefore, the old LGA1366 platform will still remain demanded after the launch of the new LGA1155 systems: it will be the only system supporting full-speed video sub-systems including multiple GPUs connected via PCIE bus with the maximum bandwidth.
- An important change has been made to the display interfaces as well. The graphics core in the new processors will support HDMI 1.4, which key peculiarity is 3D support.

5. Power control unit and Next Gen Turbo Boost Technology :

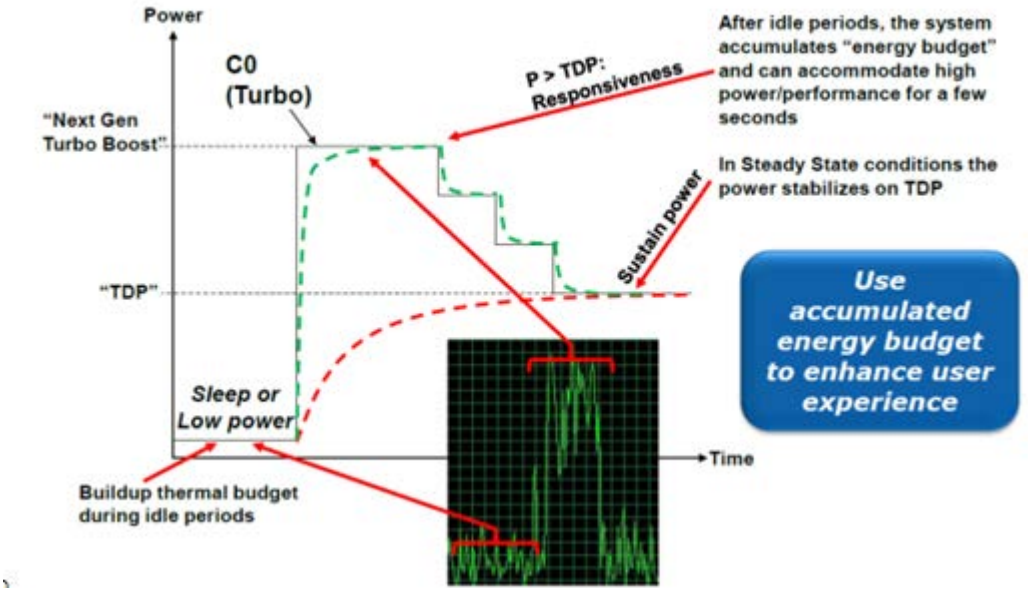
- It is a programmable micro-controller that collects information about temperatures and currents in different parts of the processor and can control its voltages and frequencies interactively.
- PCU implements power-saving functions as well as Turbo-mode.
- All functional units of the Sandy Bridge CPUs are split into three domains, each using its own independent clocking and powering algorithms.

    - Processor cores and L3 cache working at the same frequency and voltage.
    - Graphics core working at its own frequency.
    - System Agent

- This structure allowed Intel to implement Enhanced Intel SpeedStep and Turbo Boost technologies simultaneously and independently for the processor cores and the GPU.
- Sandy Bridge has a fully hardware solution that controls the frequencies of computational and graphics cores interdependently, considering their current power consumption. This way processor cores can be overclocked more effectively in Turbo mode when the graphics core is idle, and Vice versa.



- The advantages of its new implementation imply that the PCU can control the frequencies in a more intellectual way, taking into account the actual temperatures of the processor units, and not only their power consumption. It means that when the CPU works in favorable thermal conditions, its power consumption can exceed the TDP maximum.
- Processor load is usually very uneven during typical everyday work. Most of the time the CPU is in power-saving state and needs to work very fast only for short periods of time. The processors doesn't get overheated within these short periods due to inertia from the cooler heat conductivity.
- The PCU unit in Sandy Bridge processors that controls the frequencies believes that nothing bad will happen if at this time the CPU is overclocked more dramatically than the TDP can theoretically allow. As soon as the processor temperature starts approaching the critical thresholds, the frequency will be dropped down to safe levels.



- This automatically suggests that it more beneficial to use highly efficient cooling systems in Sandy Bridge based platforms for achieving maximum performance. But do not get overexcited: there is a hardware limitation that will restrict the maximum length of "beyond TDP" operation with 25 seconds.
- The main issue is that this clock frequency generator is the only one in the system and it is responsible for generating all the frequencies. And as we know, not all busses and controllers can actually tolerate overclocking that well. For example, when we increase the PCI Express frequency or USB or SATA controllers speed, instability may occur fairly quickly. This factor is going to be a serious obstacle during CPU overclocking by raising the clock generator frequency.
- 2nd Generation Core I processors overclocking :
  - Frequency knobs
    - Core frequency unlocked in 100 MHz increment Graphics frequency unlocked in 50 MHz increment
    - Memory frequency unlocked in 266 MHz increment
    - BCLK adjustable in <1MHz increment
  - OC knobs
    - Core and Graphic power limit override
    - XMP and Manual memory timing override supported
  - Independent voltage control allowed on core , Graphic and Memory
- Sandy Bridge processors have their clock generator frequency set at 100 MHz. The generator allows varying this frequency in a very broad range even with 0.1 MHz increment. However, any attempts to increase this frequency rapidly lead to system instability or failure. At this point we don't know about any successful attempts to push the clock generator frequency beyond 105 MHz. In other words, the time tested overclocking method of raising clock generator frequency doesn't work in Sandy Bridge processors and can only allow a 5% increase at best.
- Intel is going to offer several Sandy Bridge based CPU modifications that will have an unlocked frequency multiplier and should theoretically overclock to 5.7 GHz (57 is the maximum multiplier allowed by this microarchitecture). However, these processors marked with a "K" in their model no
- The multiplier adjustment will not affect Turbo Boost technology, which will also contribute to this manual frequency increase. Moreover, Intel won't limit the multipliers for the graphics core and memory frequency. In other words, it will be possible to overclock memory and graphics in any Sandy Bridge based system: a regular or an enthusiast one.

Intel Ivy Bridge Microarchitecture:

Features:

1. Tri-gate transistor ("3-D") technology (up to 50% less power consumption at the same performance level as 2-D planar transistors).
2. PCI Express 3.0 support.
3. Max CPU multiplier of 63 (57 for Sandy Bridge).
4. RAM support up to 2800 MT/s in 200 MHz increments.
5. Intel HD Graphics 2500/4000 with DirectX 11, OpenGL 3.1, and OpenCL 1.1 support.
6. The built-in GPU will have up to 16 execution units (EUs), compared to Sandy Bridge's maximum of 12.
7. A new random number generator and the RdRand instruction, codenamed Bull Mountain.
8. Intel Quick Sync Video.
9. Low voltage DDR3L for mobile processors.
10. Multiple 4K video playback.
11. Thermal design power (TDP) will come in 77/65/55/45/35 W options for desktop processors, while Intel says that the mobile processors will use a configurable TDP.

Suffixes of model no of core-I processors denote:

- K - Unlocked (adjustable CPU multiplier up to 63 bins)
- S - Performance-optimized lifestyle (low power with 65W TDP)
- T - Power-optimized lifestyle (ultra- low power with 35-45W TDP)

Performance: Compared to Sandy Bridge

1. 5% to 15% increase in CPU performance
2. 20% to 50% increase in integrated GPU performance

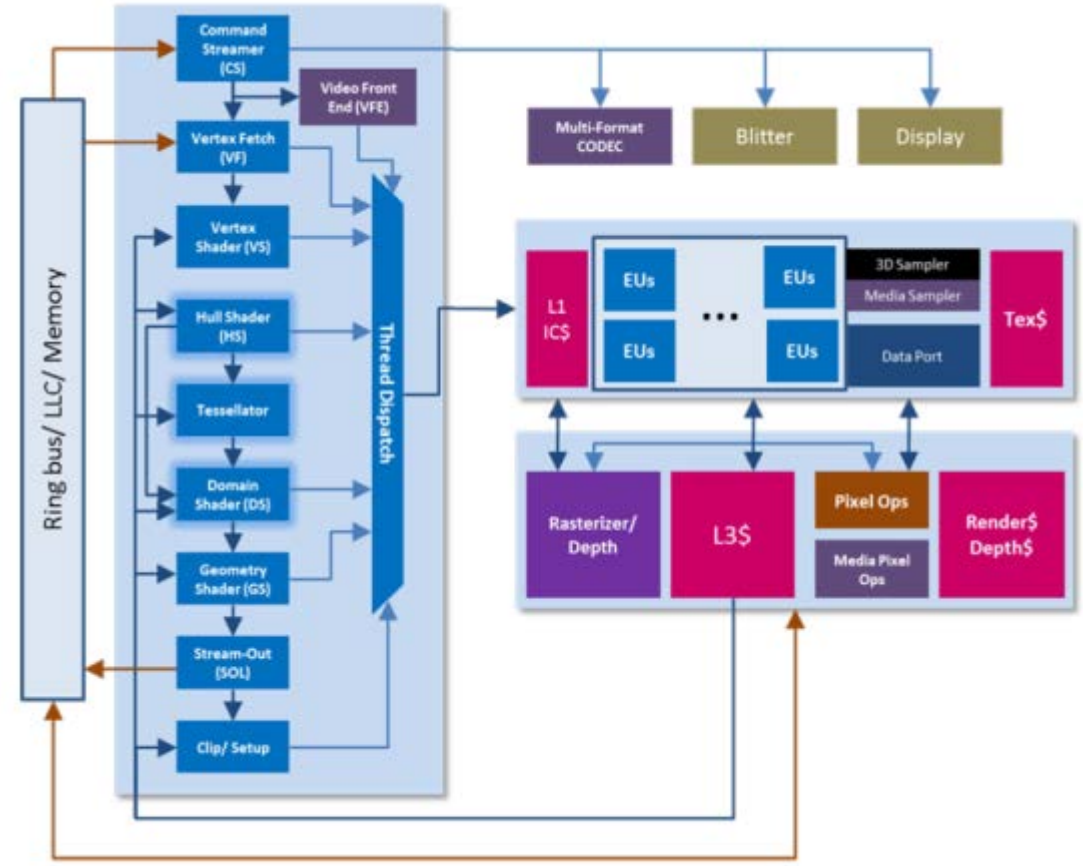
Architecture:

1. Increased Efficiency:

- The new production process will allow Intel to significantly lower the heat dissipation. And it is actually not only due to the transition to 22 nm, but also due to the involvement of Tri-Gate Transistors. These transistors allow eliminating leakage currents very effectively.
- Intel claims that the performance-per-watt is going to almost double compared with what is from Sandy Bridge.
- Intel integrated DDR3L memory support into their current memory controller. DDR3L is the memory with lower power consumption that uses only 1.35 V voltage feed. It has feature of lower signal voltages, but also about the ability to shut off I/O power to DDR memory in deep sleep states.
- But the absolute killer feature in the upcoming Ivy Bridge, which will be related to heat dissipation, will be configurable TDP and Low Power Mode.
- As you know, each currently available Intel processor has a clearly identified TDP, within which the frequency may sometimes be increased with the help of Turbo Boost technology. Ivy Bridge processors will have three TDP options for each CPU model: the minimum TDP, the nominal TDP and the maximum TDP. It means that with proper cooling and sufficient power supply the CPU can significantly exceed its nominal frequency without any concerns for the limitations of the nominal TDP..
- It will certainly work in parallel with Enhanced Intel SpeedStep and Turbo Boost. The major difference in this case is that EIST and Turbo Boost are independent technologies, while configurable TDP works in manual mode only. User will be also to switch the TDP modes using a special switch on the laptop, or it will be done via special proprietary software developed and supplied by the notebook makers themselves.
- According to the developers, in this mode the processor will try to save as much power as possible by distributing all computational threads among as few cores as they can in order to switch as many cores as possible into power-saving mode. In regular operational modes the computational threads are evenly distributed among the cores to improve the system response time.

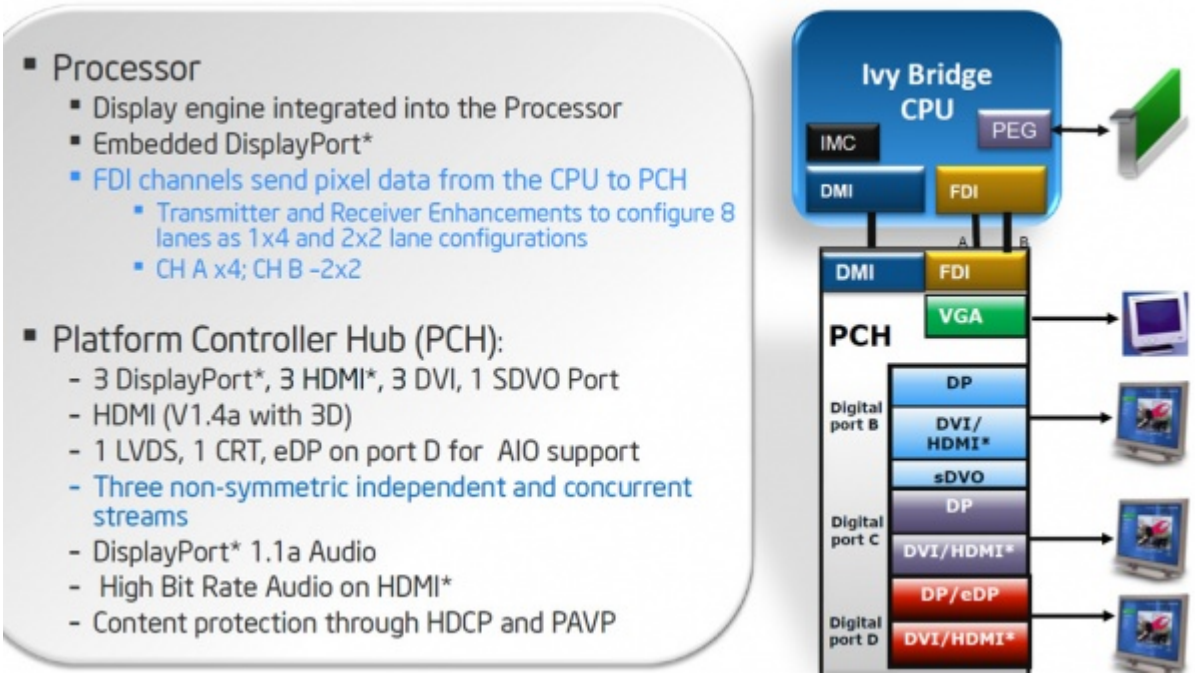
2. Intel HD Graphics 4000

➢ HD Graphic Microarchitecture :



➢ Enhancement from Sandy Bridge:

- Scalable Architecture partitioned into 5 domains :
    - Global Assets : Includes Geometry Front end up to Setup
    - Slice Common : Includes Rasterizer L3 cache and Pixel Back-end
    - Slice : shaders (EUs), ICs, Samplers ,Address Generator
    - CODEX and media
    - Display
- Set the stage for further scale up opportunity
- Improved Geometry Performance
    - Faster GS and H/w Stream out
    - Faster Clip/Setup
- Fast clear of Render Target
- Sampler throughput (peak GFLOPs)
    - Increased no. of threads/registers to cover latency and support complex shader
    - Enhanced coissue
- L3 cache lowers BW need from Ring Architecture
- Pretty much all of the above domains have been improved on in some way including the addition of tessellation units, vastly improved anisotropic quality and a higher possible thread count for complex shaders.



- An updated FDI (flexible display input) allows for three concurrent outputs for some impressive integrated graphics multi-monitor solutions.

Category	Feature	HD 4000 (Ivy Bridge)	HD 2000/3000 (Sandy Bridge)
Architecture Improvement	Unified Shader Arch	Yes	Yes
	Execution Units (EUs)	6/16 EUs	6/12 EUs
	Dedicated Math box	Yes	Yes
	Media Processing	Yes	Yes
	Targeted OS optimization	Win7/Win8	Win7/Vista/XP
3D performance	Core Frequency	Up to 1350 MHz	Up to 1350 MHz
	DirectX support	DX11	DX10.1
	OPEN GL support	Open GL 3.1	Open GL 3.0
	Shader Model support	SM 4.1	SM 4.1
	Dynamic Frequency Scaling	Yes	Yes
	Maximum Resolution	2560*1600	2560*1600
	HDMI (V 1.4 with 3D) support	Yes	Yes

- This table gives us a quick top-down view of the changes moving from Sandy Bridge to Ivy Bridge for the integrated GPU. Clock speeds still remain the same with an "up to" level of 1350 MHz depending on the Turbo Boost and shared TDP limits at each instant. Newer version of OpenGL and DirectX are supported so the HD 4000 graphics should at the very least be more compatible with games and applications on the market.

3. Overclocking To Get Better, but Just a Little

- In reality overclockers' biggest issue with the LGA1155 platform is that there is no straight-forward way to overclock anything by simply adjusting the BCLK frequency.
- Intel didn't give up on overclockers and introduced four pretty useful features.

i. The maximum multiplier for the new Ivy Bridge processors will be x63 instead of x57, as it used to be. As a result it is possible to achieve much higher frequencies. But nevertheless, we are not talking about anything like 8.4 GHz, which AMD Bulldozer turned out capable of.
ii. Intel is going to make their memory controller support much faster DDR3 SDRAM. The fastest DDR3 SDRAM Sandy Bridge currently supports is DDR3-2133 memory, but they promise us we will be able to use up to DDR3-2800.
iii. There is a smaller adjustment increment for the memory frequency than the 266 MHz that we have today. According to Intel, it should be no more than 200 MHz
iv. The most interesting thing is the ability to dynamically adjust the processor clock frequency multiplier without a system reboot. It means that there will appear multiple utilities, which will allow not only overclocking the processor easily, but also using some unique algorithms to adjust the processor frequency.

4. Key ISA visible changes :

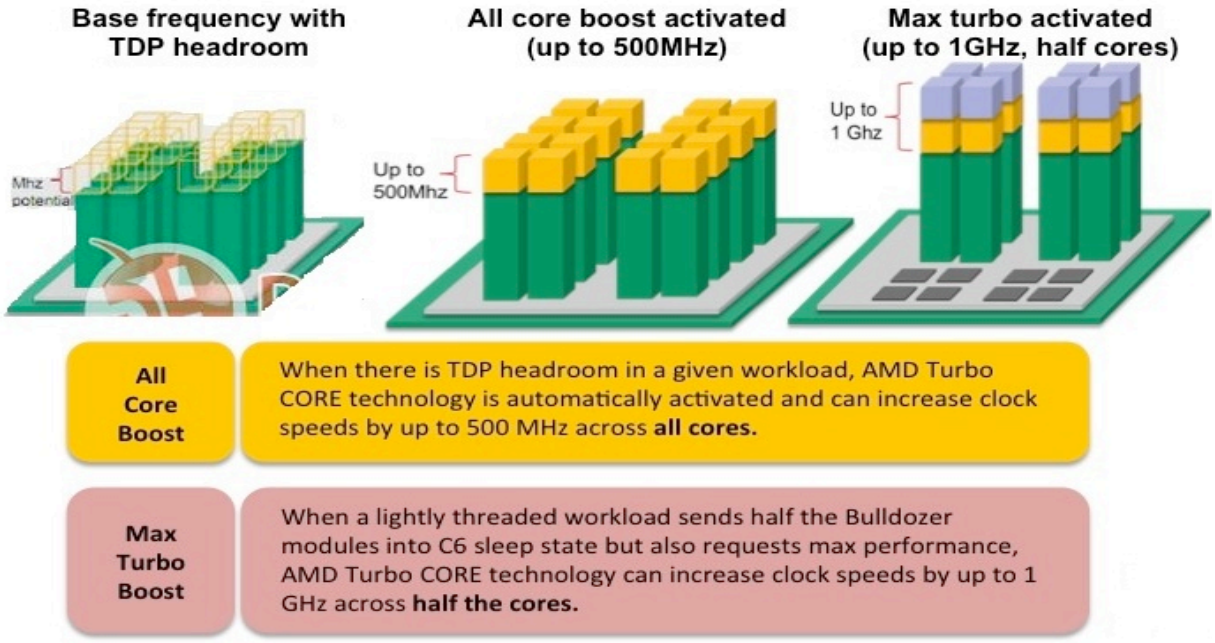
- Digital Random Number Generator
- Supervisor Mode Execution Protection (SMEP)
  - prevent an alien application from invading the OS services with higher privileges
- REP MOVSB/STOSB performance improvements
  - More consistent performance along string length
- Fast access of FS and GS registers
  - Useful for user level thread storage by providing 4 new instructions for ring -3 access of FS and GS register
- Float16 format conversion instruction
  - Conversion between 16 bit (compressed) Floating point memory format and 32 bit Single precision (256 bit AVX and 128 bit SSE versions)

3. PROCESSOR SPEEDUP TECHNOLOGIES OF AMD

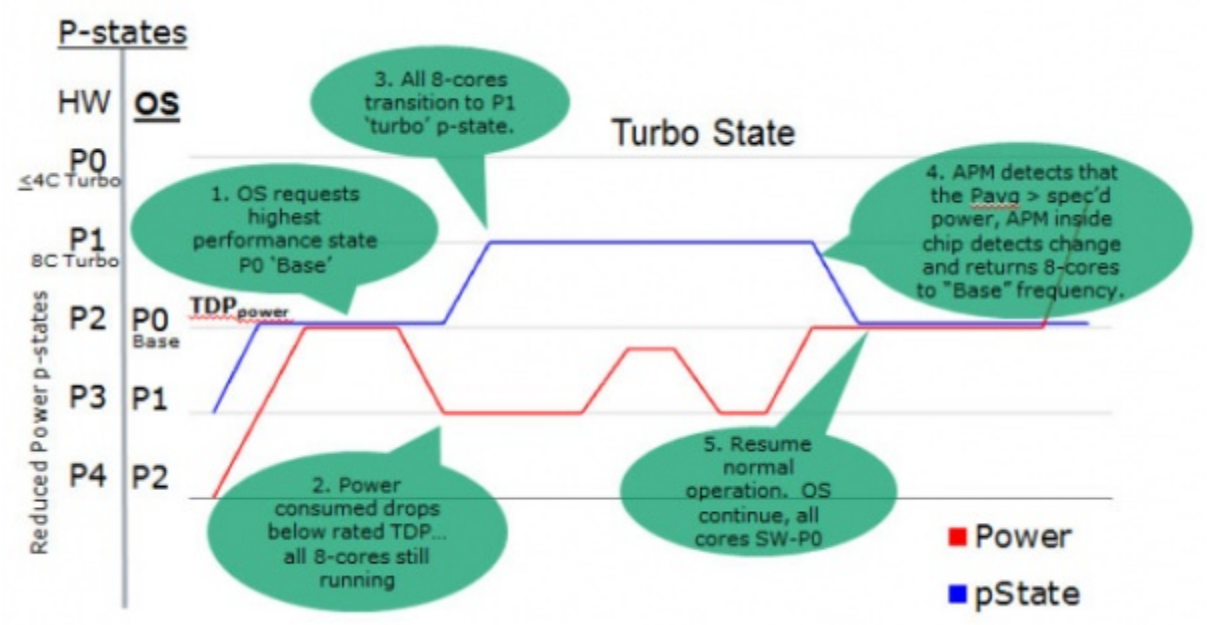
3.1) AMD TURBO CORE TECHNOLOGY

- The processor is essentially divided into three parts. The Uncore is clocked at its usual speed and does not change or power down.
- The six cores are essentially divided into two processor blocks of three a piece.
- When the processor determines that three or less cores are being fully utilized, it will place the underutilized cores into a lower power, low clock peed state. In this case the cores will be clocked at 800 MHz and will have their voltage decreased. The other three cores will be placed into a boost mode, with their clock speeds increased as well as their voltages per core increased. AMD is designing these cores to be run between 400 MHz and 500 MHz faster. A theoretical six core clocked at 3.2 GHz would then see a boost of three cores up to 3.6 GHz.
- By downclocking the other cores and cutting their voltage, AMD is able to keep the processor in essentially the same TDP envelope.
- AMD does not have the ability to totally power down and shut off those cores, unlike Intel.
- The big key here to take away is that this process is invisible to the OS and applications that the CPU is running. It is a totally controlled in hardware, so at most a BIOS update is needed for motherboards.

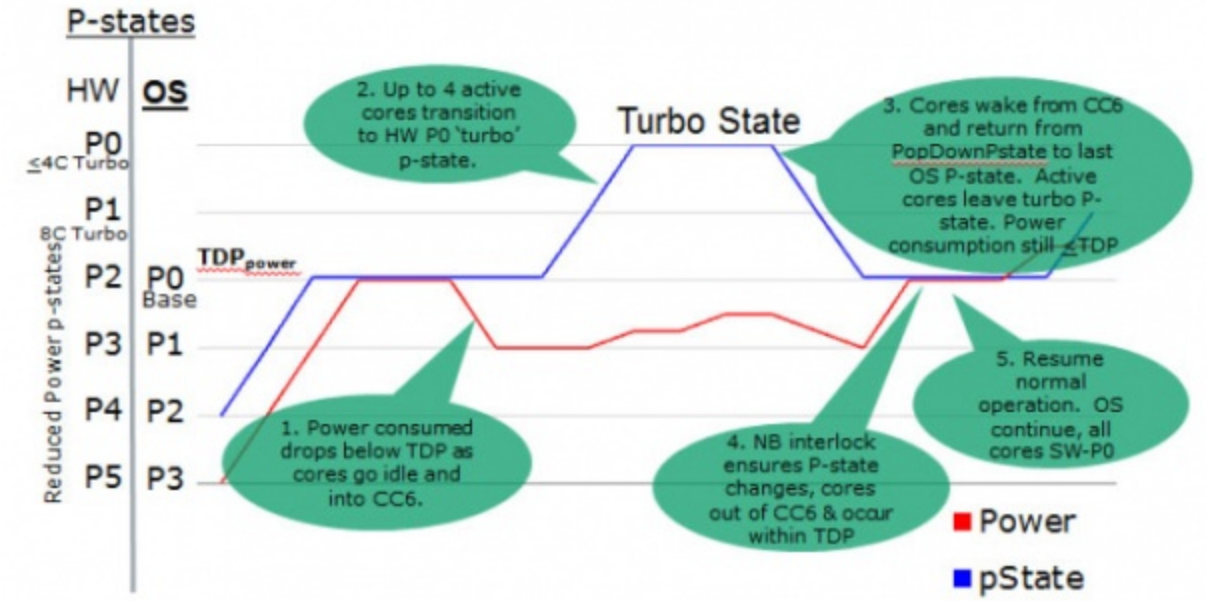
New AMD TURBO CORE TECHNOLOGY



- When the FX processor detects that it has available TDP headroom, it can temporarily increase the clock speed (and thus the power draw) of the CPU in order to improve performance. The frequency can go up a modest amount when even all 8 cores of the FX CPU are in use (a change from the previous iteration of Turbo Core) and go even higher when 4 or fewer cores are in use.



- The "Max Core" mode as it is called allows our FX-8150 to jump from the 3.6 GHz base frequency to a speed of 3.9 GHz for a short period of time even when all 8 cores are in use. In the diagram above the blue bar's highest position is that 3.9 GHz setting and it is important to note that the time period that the CPU can run at that speed will vary based on the individual CPU, the cooling solution at work and some other indicators.
- Bulldozer includes an integrated "AMD Power Manager" that monitors these conditions and tells the CPU when it should return back to the standard P0 base setting and clock speed.



- For lightly threaded apps, specifically those times when 4 or fewer cores are in use, the AMD FX processors can jump to a higher frequency of Turbo but will still eventually drop to the base state after some period of time.
- Obviously the advantages of this technology from AMD are the same as they are from Intel: higher performance on single and dual-threaded applications than you would otherwise be able to achieve by taking advantages of the TDP headroom of a large die when only a portion of it is being utilized.

Processor Model	CPU Base	CPU Turbo Core	CPU Max Turbo	TDP	Cores	L2 Cache	L3 Cache	Max DDR3	PKG	NB
FX-8150*	3.6GHz	3.9GHz	4.2GHz	125W	8	8MB	8MB	1866	AM3+	2.2GHz
FX-8120	3.1GHz	3.4GHz	4.0GHz	125W						2.2GHz
FX-6100	3.3GHz	3.6GHz	3.9GHz	95W	6	6MB				2.0GHz

3.2) AMD Balanced Smart Cache

☐ Shared L3 cache (either 6MB or 4MB)
☐ 512KL2 cache per core

- Benefit: Shortened access times to the highly accessed data for better performance

3.3) Integrated DRAM Controller with AMD Memory Optimizer Technology

- A high-bandwidth, low-latency integrated memory controller
- Supports PC2-8500 (DDR2-1066); PC2-6400 (DDR2-800), PC2-5300 (DDR2-667), PC2-4200 (DDR2-533) or PC2-3200 (DDR2-400) SDRAM unbuffered DIMMs – AM2+
- Support for unregistered DIMMs up to PC2 8500(DDR2-1066MHz) and PC3 10600 (DDR3-1333MHz) – AM3
- Up to 17.1GB/s memory bandwidth for DDR2 and up to 21GB/s memory bandwidth for DDR3
    - Benefit: Quick access to system memory for better performance.

3.4) AMD's HyperTransport™ Technology

Introduction

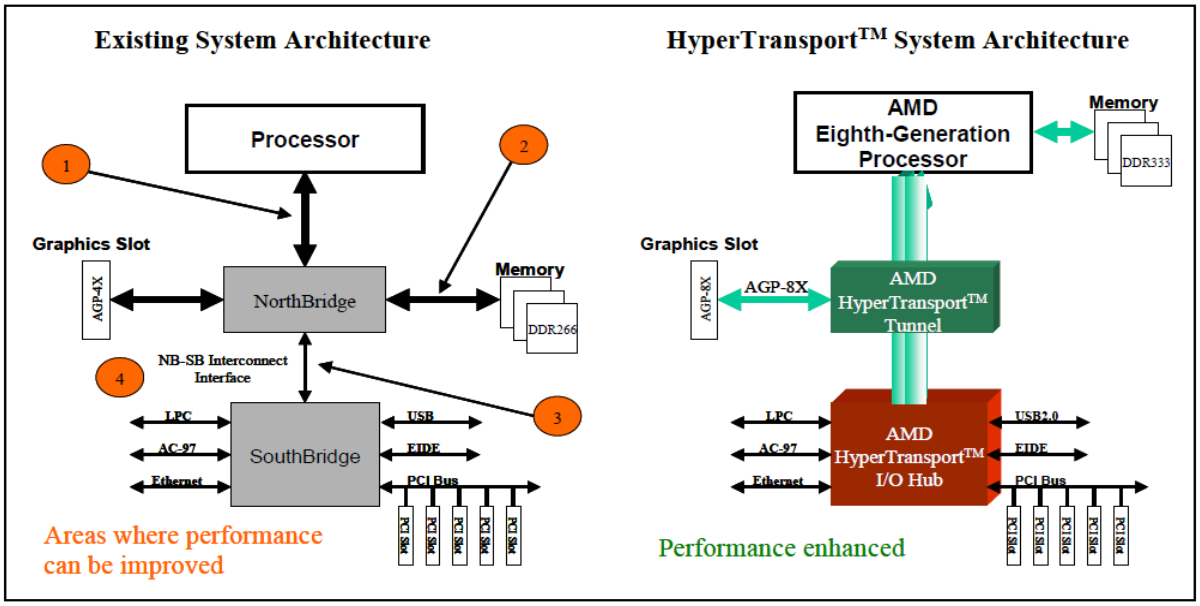
- Incorporating HyperTransport technology as the chip-to-chip interconnect, these core logic elements integrate the latest I/O technologies to form the system building blocks for a wide range of high-performance platforms harnessing the power of AMD's eighth-generation AMD Athlon processor for mobile and desktop applications and AMD Opteron processor for workstation and server applications.
- HyperTransport technology is a high-speed, high-performance, point-to-point link for integrated circuits, and is designed to meet the bandwidth needs of tomorrow's computing and communications platforms.
- At a peak throughput of up to 12.8GB/s per link, HyperTransport technology provides an I/O solution for the most demanding system applications.

System Performance Enhancement

HyperTransport technology is designed to increase overall performance by helping to remove I/O bottlenecks, which improves bandwidth and reduces latency.

- The processor front-side bus
- Memory interface
- Chip-to-chip interconnect
- I/O expansion capability to high-speed industry buses

1. **Performance Enhancement 1: Processor Front-Side Bus**
    - For optimal performance, the front-side bus bandwidth must scale with increasing processor speeds. Current front-side bus bandwidth on AMD's seventh generation platforms is on the order of 2.1GB/s.
    - Replacing what has traditionally been the system "front-side bus" with a HyperTransport technology-based I/O connection dramatically extends processor to system communication bandwidth from 2.1GB/s up to 6.4GB/s (and potentially 12.8GB/s with future devices).



System Architecture

2. Performance Enhancement 2: Memory Interface

- When cache misses occur, the processor must fetch information from main memory. In the Northbridge/Southbridge architecture, memory transactions must traverse through the Northbridge element, creating additional latencies that reduce performance potential.

- To help resolve this performance bottleneck, AMD incorporates the memory controller into its eighth-generation processor. The direct interface to the memory can significantly reduce the memory latency seen by the processor.
- This latency will continue to drop as the processor frequency scales. Additionally, hardware and software memory prefetching mechanisms can further reduce the effective memory latency seen by the processor.

3. Performance Enhancement 3: Chip-to-Chip Interconnect

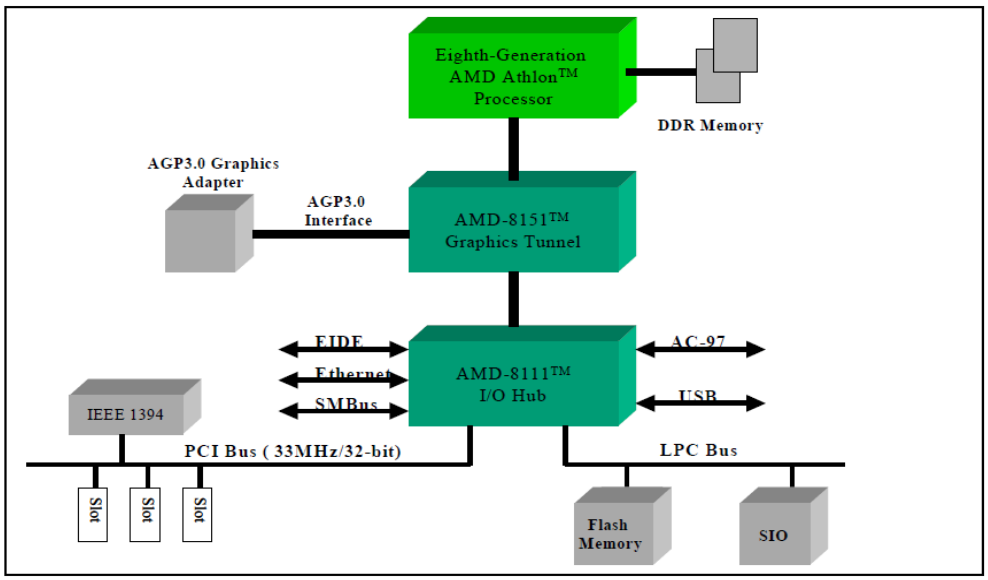
- Current interface schemes offer throughput performances on the order of 266MB/s to 1GB/s. These rates may be sufficient for desktop platforms; however, a more robust interface is required for workstation, server, and other future platforms.
- The simultaneous integration of high-speed technologies , onto the high-end platforms will quickly dwarf the bandwidth capabilities of the existing interfaces.
- HyperTransport technology provides a high-speed, chip-to-chip interconnect that virtually eliminates the I/O performance bottleneck with ample performance headroom for future growth.

4. Performance Enhancement 4: I/O Expansion Capability to High-Speed Industry Buses

- The functionality would have to be attached to an existing bus interface such as the PCI bus. However, an existing bus may not have sufficient bandwidth to support high-speed technologies, especially in instances where multiple buses or combinations of buses must be supported simultaneously.
- The functionality would have to be directly attached to the higher-speed proprietary chip-to-chip interconnect bus via a bridging device. However, the proprietary nature of this solution may limit the number of components available from vendors, thus impacting cost and availability.
- The functionality would have to be integrated into one of the core logic components. This solution is the least flexible, as a wide range of components would have to be created for each desired combination of feature-set buses.

HyperTransport Technology-Based System Architecture

- HyperTransport technology-based system architecture represents a significant enhancement to traditional system architectures by providing daisy-chained interconnects between system components.
- The concepts of Northbridge, Southbridge, and PCI-bus chip-to-chip Interconnects are now replaced with tunnels, I/O hubs, HyperTransport technology links, and HyperTransport technology I/O chains.
- Systems based upon HyperTransport technology are no longer limited to a two-chip core-logic solution. Multiple components can be chained together in a building-block fashion to produce a multitude of platform configurations.
- HyperTransport technology has ample bandwidth to support the various high-speed bus technologies, even in combination implementations. The result is an extremely high-performance system architecture that is flexible, adaptive, and reliable.



High Performance AMD HyperTransport Technology System Architecture

3.5) AMD 3DNOW!

- 3DNow! is an extension to the x86 instruction set developed by Advanced Micro Devices (AMD).
- It adds single instruction multiple data (SIMD) instructions to the base x86 instruction set, enabling it to perform simple vector processing, which improves the performance of many graphic-intensive applications.
- The first microprocessor to implement 3DNow! was the AMD K6-2, which was introduced in 1998. When the application was appropriate this raised the speed by about 2-4 times.
- As an enhancement to the MMX instruction set, the 3DNow! Instruction-set augmented the MMX SIMD registers to support common arithmetic operations (add/subtract/multiply) on single-precision (32-bit) floating-point data.

- The first implementation of 3DNow! technology contains 21 new instructions that support SIMD floating-point operations. The 3DNow! data format is packed, single-precision, floating-point. The 3DNow! instruction set also includes operations for SIMD integer operations, data prefetch, and faster MMX-to-floating-point switching. Later, Intel would add similar (but incompatible) instructions to the Pentium III, known as SSE for Streaming SIMD Extensions.

- **3DNow! extensions**
  - This extension to the 3DNow! instruction set was introduced with the first-generation Athlon processors. The Athlon added 5 new 3DNow! instructions and 19 new MMX instructions. Later, the K6-2+ and K6-III+ (both targeted at the mobile market) included the 5 new 3DNow! instructions, leaving out the 19 new MMX instructions. The new 3DNow! instructions were added to boost DSP. The new MMX instructions were added to boost streaming media.

- **Who is best 3DNOW or MMX?**
  - The 19 new MMX instructions are a subset of Intel's SSE1 instruction set. In AMD technical manuals, AMD segregates these instructions apart from the 3DNow! extensions. In AMD customer product literature, however, this segregation is less clear where the benefits of all 24 new instructions are credited to enhanced 3DNow! technology.
  - This has led programmers to come up with their own name for the 19 new MMX instructions. The most common appears to be Integer SSE (ISSE).SSEMMX and MMX2 are also found in video filter documentation from the public domain sector.

- 3DNow! Professional
  - 3DNow! Professional is a trade name used to indicate processors that combine 3DNow! technology with a complete SSE instructions set (such as SSE1, SSE2 or SSE3). The Athlon XP was the first processor to carry the 3DNow! Professional trade name, and was the first product in the Athlon family to support the complete SSE1 instruction set (for the total of: 21 original 3DNow! instructions; 5 3DNow! extension DSP instructions; 19 MMX extension instructions; and 52 additional SSE instructions for complete SSE1 compatibility).

- 3DNow! and the Geode GX/LX
  - Two instruction which now is absent in all the processors.

- **Advantages and disadvantages**
  - One advantage of 3DNow! is that it is possible to add or multiply the two numbers that are stored in the same register. With SSE, each number can only be combined with a number in the same position in another register. This capability, known as horizontal in Intel terminology, was the major addition to the SSE3 instruction set.
  - A disadvantage with 3DNow! is that 3DNow instructions and MMX instructions share the same register-file, whereas SSE adds 8 new independent registers (XMM0 - XMM7.)
  - Because MMX/3DNow! registers are shared by the standard x87 FPU, 3DNow! instructions and x87 instructions cannot be executed simultaneously. However, because it is aliased to the x87 FPU, the 3DNow! & MMX register states can be saved and restored by the traditional x87 F(N)SAVE and F(N)RSTOR instructions. This arrangement allowed operating systems to support 3DNow! with no explicit modifications, whereas SSE registers required explicit operating system support to properly save and restore the new XMM registers (via the added FXSAVE and FXRSTOR instructions.)
  - The FX* instructions are an upgrade to the older x87 save and restore instructions because these could save not only SSE register states but also those x87 register states (hence which meant that it could save MMX and 3DNow! registers too).
  - On AMD Athlon XP and K8-based cores (i.e. Athlon 64), assembly programmers have noted that it is possible to combine 3DNow! and SSE instructions to reduce register pressure, but in practice it is difficult to improve performance due to the instructions executing on shared functional units.

3.6) AMD PowerNow!™ Technology

- PowerNow! is speed throttling and power saving technology of AMD's processors used in laptops.
- The CPU's clock speed and VCore are automatically decreased when the computer is under low load or idle, to save battery power, reduce heat and noise.
- The lifetime of the CPU is also extended because of reduced electromigration, which varies exponentially with temperature.
- The technology is a concept similar to Intel's SpeedStep technology.

Advantages:

- Optimizes battery life
- Provides performance on demand when required by the application
- Allows the processor to dissipate less heat under normal operating conditions, providing a cooler and quieter-running notebook
- Operates automatically in the background

3.7) Cool 'N' Quiet™ 2.0 Technology

Cool'n'Quiet™ 2.0 Technology reduces heat and noise so you can experience amazing performance without distraction. Combined with core enhancements, included in the AMD Phenom™ processor, that can improve overall power savings, deliver seamless multitasking and energy efficiency.

New features:

1) **Independent Dynamic Core Technology**- Helps users get more efficient performance by dynamically adjusting individual core frequencies as required by utilization needs
2) **Dual Dynamic Power Management™** - Helps improve platform efficiency by providing full-speed memory performance while enabling decreased system power consumption.
3) **AMD CoolCore™ Technology** - Helps users get more efficient performance by dynamically activating or turning off parts of the processor.
4) **AMD Wideband Frequency Control** - Allows the processor to respond more precisely to user demands, maximizing performance to deliver a better PC user experience.
5) **Multi-Point Thermal Control** - Prevents processor from creating too much heat and enables a cooler, quieter PC experience

Cool 'N' Quiet™ 3.0 Technology: **In addition to the features included with Cool'n'Quiet 2.0 Technology, the following new features have been added:**

- **AMD Smart Fetch Technology** - Fewer processing cycles are required to locate information since data storage is streamlined and stored in the shared L3 cache. Provides CPU power savings by maintaining processor sleep states and sharing cached data between cores.
- **45 nm Process Technology with Immersion Lithography** - puts more transistors in less space and delivers better processor performance while using less power.

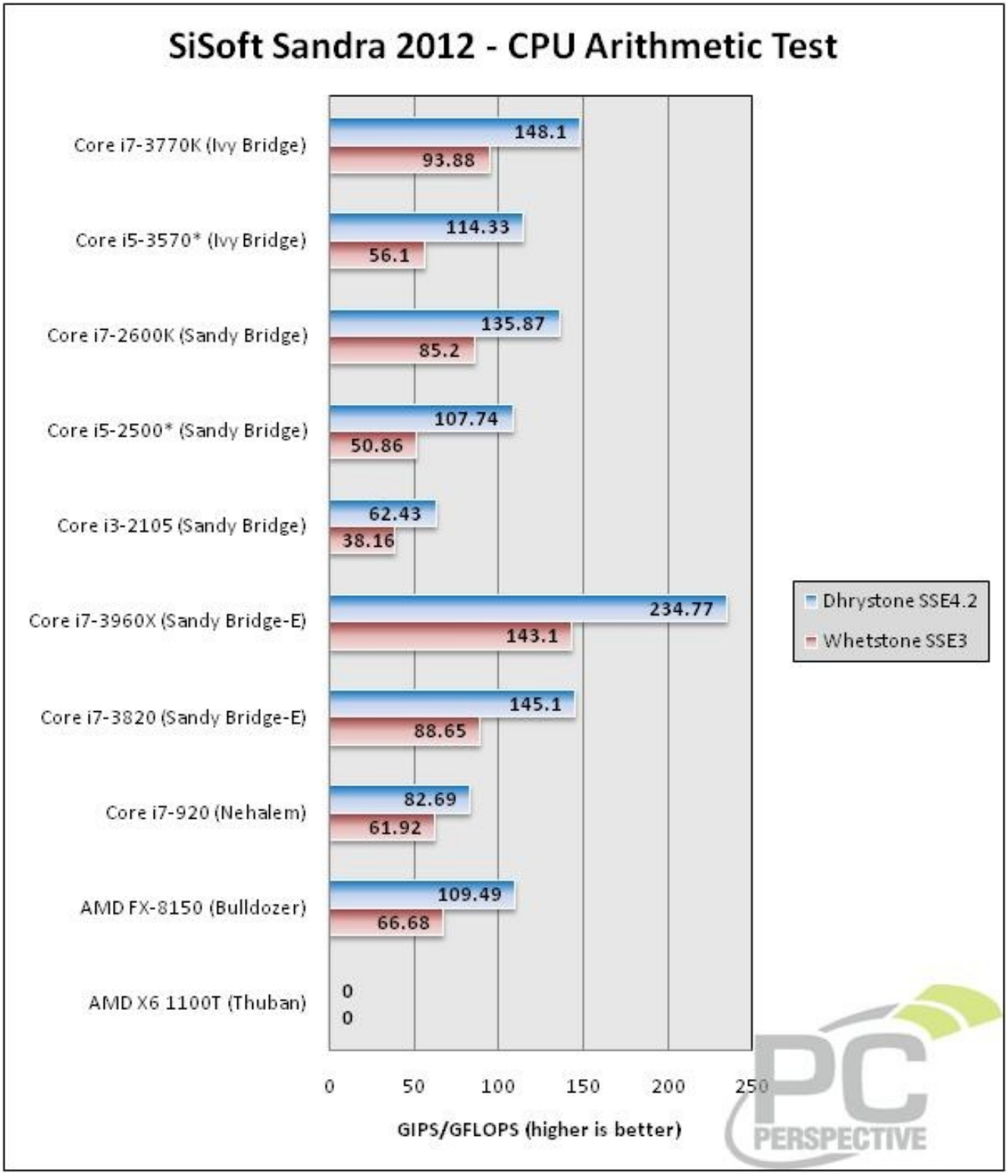
3.8) AMD CoolCore™ Technology

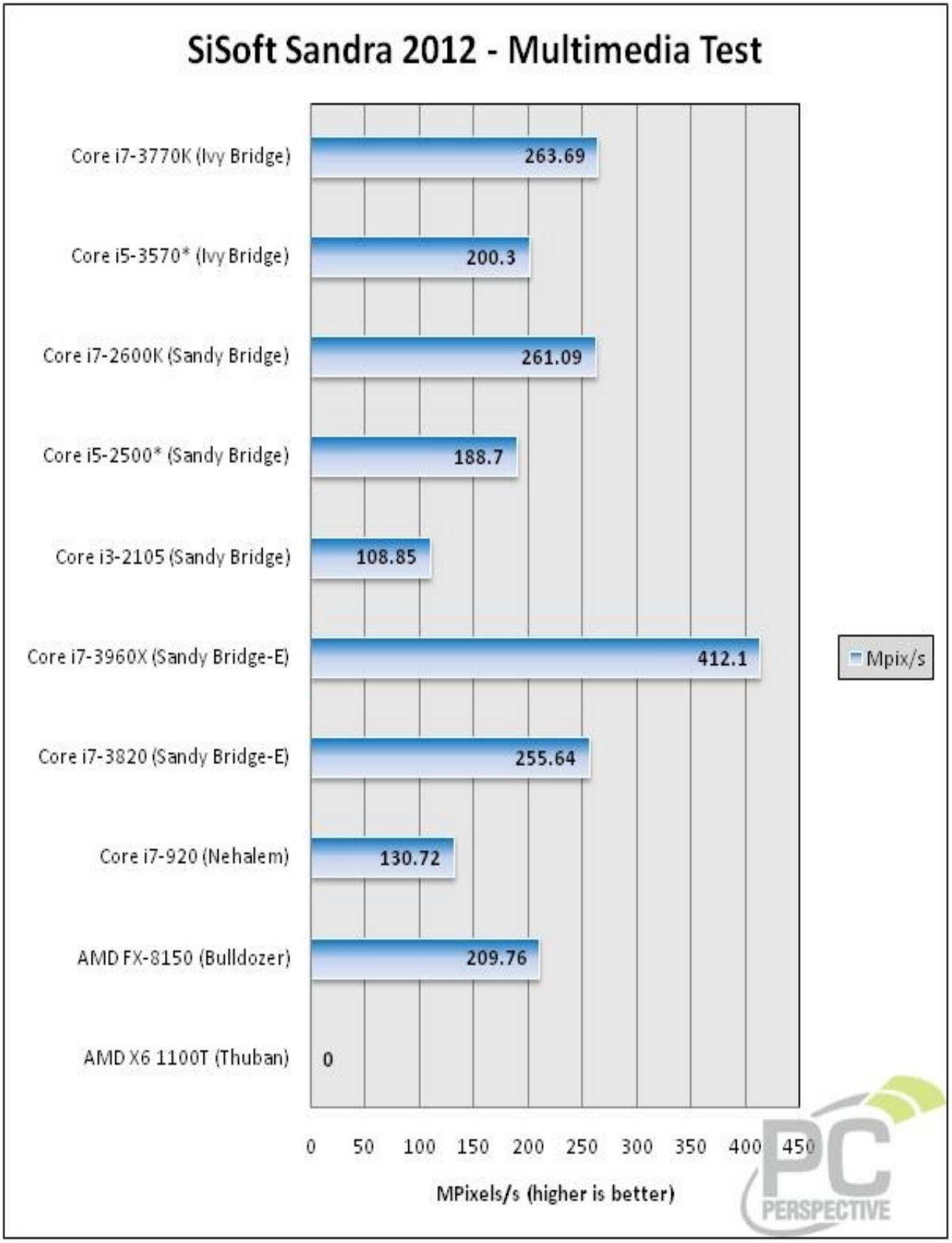
- Reduces processor energy consumption by turning off unused parts of the processor. For example, the memory controller can turn off the write logic when reading from memory, helping reduce system power.
- Works automatically without the need for drivers or BIOS enablement.
- Power can be switched on or off within a single clock cycle, saving energy with no impact to performance.
    - Benefit: Helps users get more efficient performance by dynamically activating or turning off parts of the processor.

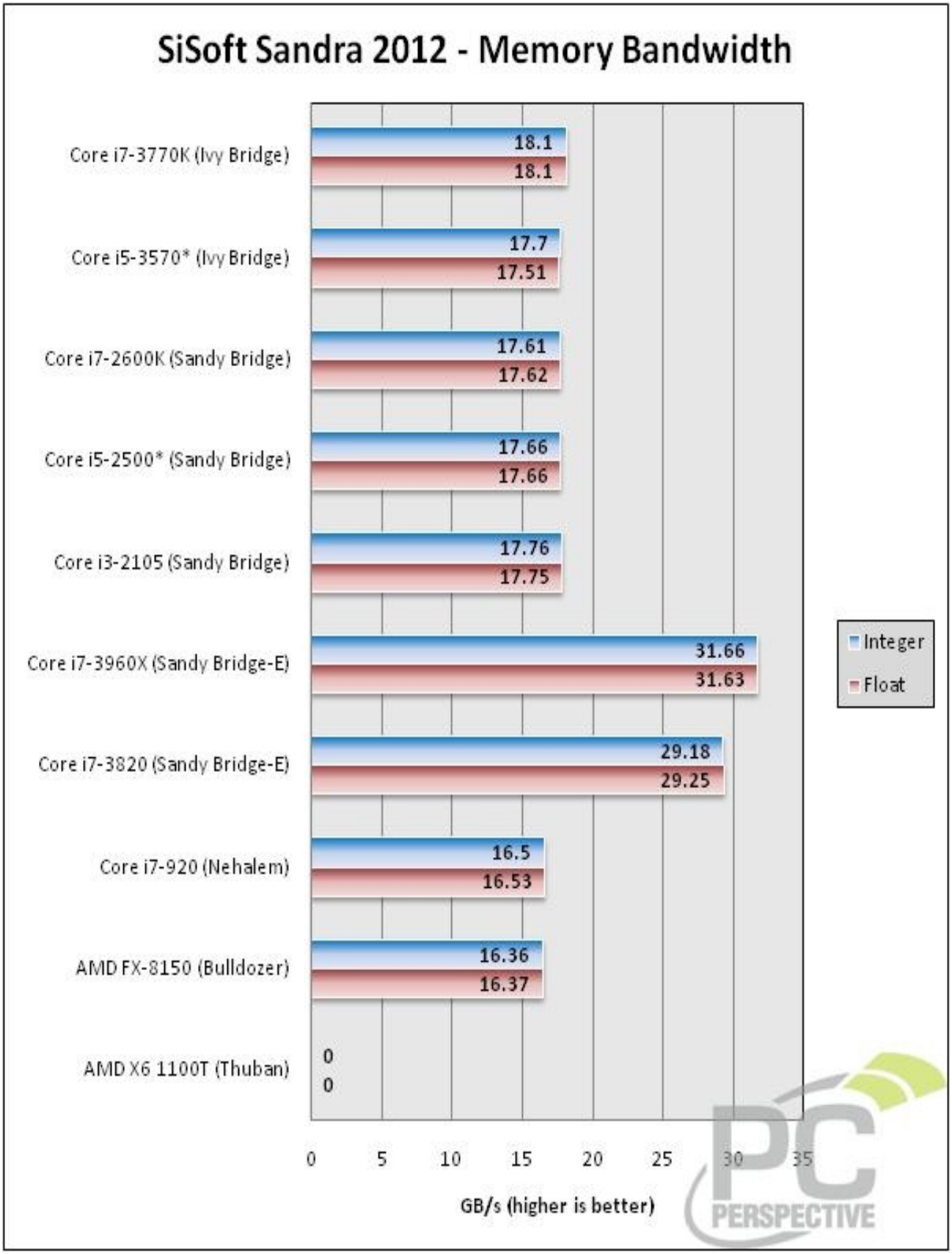
4. BENCHMARK PERFORMANCE OF VARIOUS PROCESSORS:

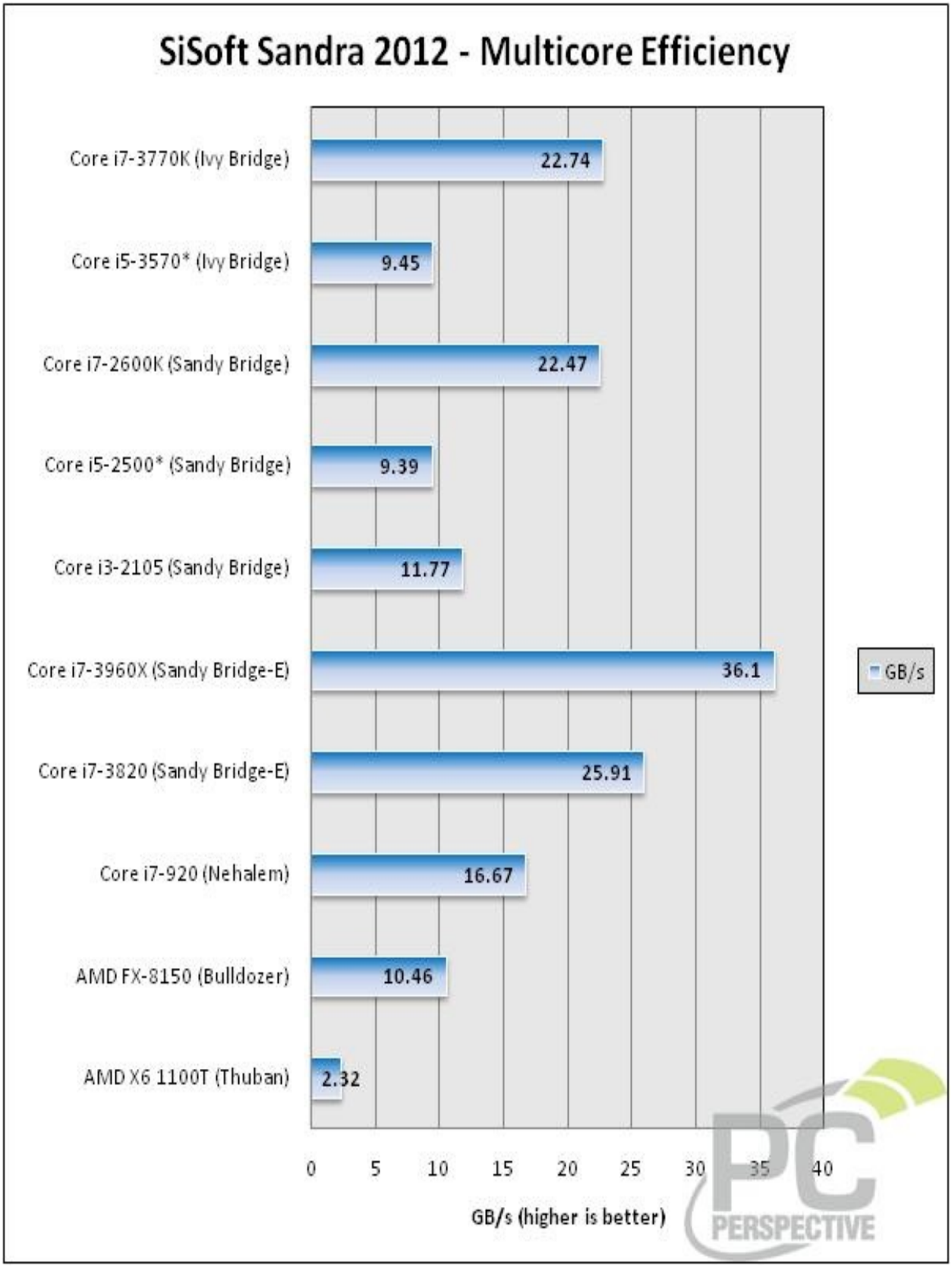
a) Comparison of various microarchitectures:

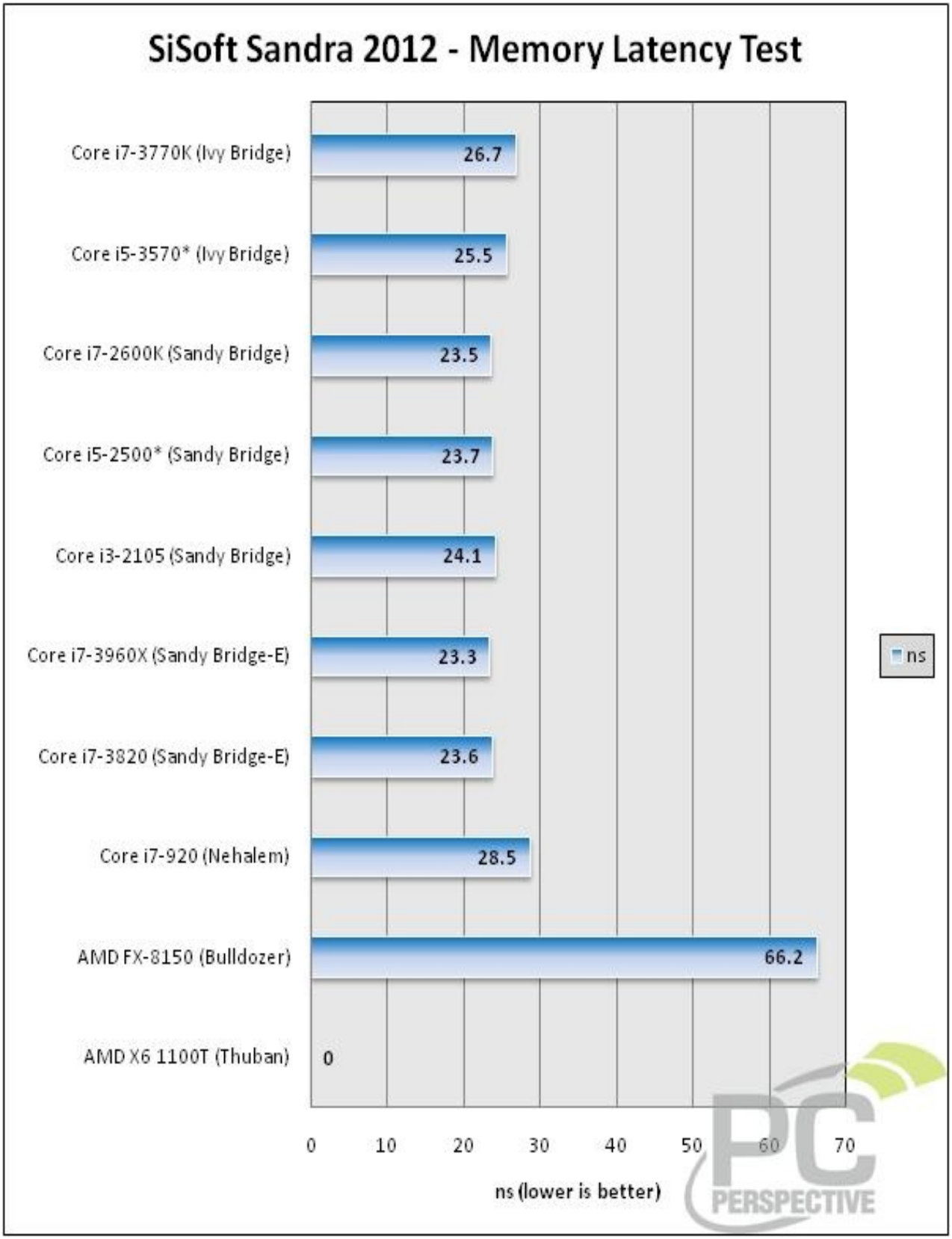
SiSoft Sandra 2012 software allows us to compare various processors for their performance.

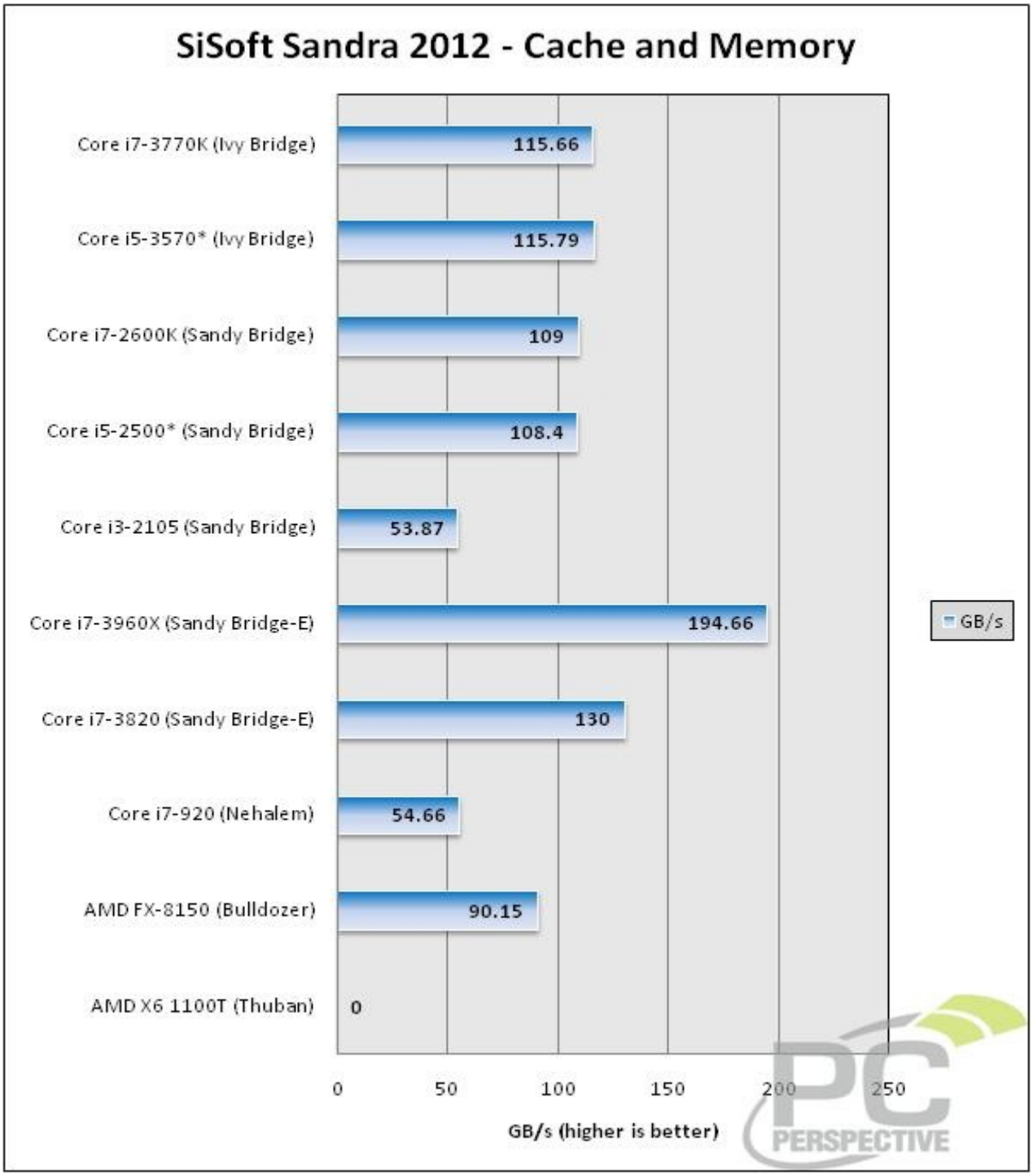



SiSoft Sandra 2012 - Multimedia Test
Core i7-3770K (Ivy Bridge)
263.69
Core i5-3570* (Ivy Bridge)
200.3
Core i7-2600K (Sandy Bridge)
261.09
Core i5-2500* (Sandy Bridge)
188.7
Core i3-2105 (Sandy Bridge)
108.85
Core i7-3960X (Sandy Bridge-E)
412.1
Core i7-3820 (Sandy Bridge-E)
255.64
Core i7-920 (Nehalem)
130.72
AMD FX-8150 (Bulldozer)
209.76
AMD X6 1100T (Thuban)
0
Mpix/s
0 50 100 150 200 250 300 350 400 450
MPixels/s (higher is better)
PC PERSPECTIVE


SiSoft Sandra 2012 - Memory Bandwidth
Core i7-3770K (Ivy Bridge)
18.1
18.1
Core i5-3570* (Ivy Bridge)
17.7
17.51
Core i7-2600K (Sandy Bridge)
17.61
17.62
Core i5-2500* (Sandy Bridge)
17.66
17.66
Core i3-2105 (Sandy Bridge)
17.76
17.75
Core i7-3960X (Sandy Bridge-E)
31.66
31.63
Core i7-3820 (Sandy Bridge-E)
29.18
29.25
Core i7-920 (Nehalem)
16.5
16.53
AMD FX-8150 (Bulldozer)
16.36
16.37
AMD X6 1100T (Thuban)
0
0
Integer
Float
0
5
10
15
20
25
30
35
GB/s (higher is better)
PC PERSPECTIVE


SiSoft Sandra 2012 - Multicore Efficiency
Core i7-3770K (Ivy Bridge)
22.74
Core i5-3570* (Ivy Bridge)
9.45
Core i7-2600K (Sandy Bridge)
22.47
Core i5-2500* (Sandy Bridge)
9.39
Core i3-2105 (Sandy Bridge)
11.77
Core i7-3960X (Sandy Bridge-E)
36.1
Core i7-3820 (Sandy Bridge-E)
25.91
Core i7-920 (Nehalem)
16.67
AMD FX-8150 (Bulldozer)
10.46
AMD X6 1100T (Thuban)
2.32
GB/s
0
5
10
15
20
25
30
35
40
GB/s (higher is better)
PC PERSPECTIVE


SiSoft Sandra 2012 - Memory Latency Test
Core i7-3770K (Ivy Bridge)
26.7
Core i5-3570* (Ivy Bridge)
25.5
Core i7-2600K (Sandy Bridge)
23.5
Core i5-2500* (Sandy Bridge)
23.7
Core i3-2105 (Sandy Bridge)
24.1
Core i7-3960X (Sandy Bridge-E)
23.3
Core i7-3820 (Sandy Bridge-E)
23.6
Core i7-920 (Nehalem)
28.5
AMD FX-8150 (Bulldozer)
66.2
AMD X6 1100T (Thuban)
0
ns
0
10
20
30
40
50
60
70
ns (lower is better)
PC PERSPECTIVE



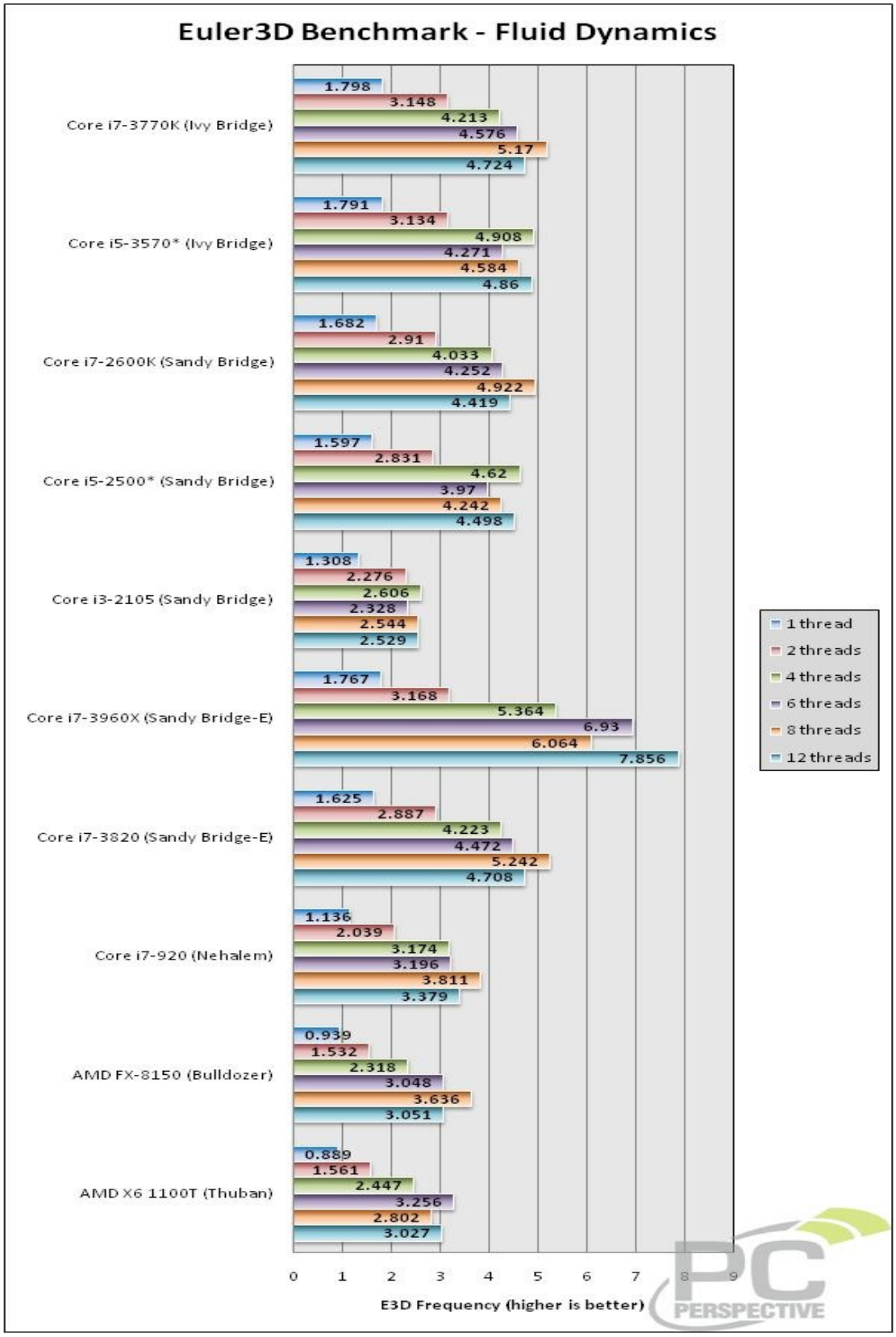
Conclusion of the comparison:

1. In the basic arithmetic test, the new Core i7-3770K is about 10% faster than the Core i7-2600K and just about matches the Core i7-3820 SNB-E part.
2. The 6-core / 12-thread beast that is the Core i7-3960X is still the top part here by a noticeable margin.
3. At the standard 1333 MHz memory clock the SNB and IVB parts seem to line up in the 17-18 GB/s range behind the SNB-E parts hitting nearly 30 GB/s.

b) Comparing various technologies:

I. Hyper threading Performance:

This fluid dynamics simulation is very CPU and memory intensive.



The Core i7-3770K is the fastest single threaded CPU tested but falls behind the Core i5-3570 at the four thread level due to the inefficiencies of Hyper Threading. Turn that thread count up to six or eight though and you see the benefits the technology can provide.

II. Intel Turbo Boost Vs. AMD Turbo Core

Intel Turbo Boost	AMD Turbo Core
Flexible policy for frequency Scaling depending op only individual cores	Fixed policy of frequency scaling depending upon set of cores
Can fully shut down core when needed	Can not fully shut down core when needed
Large increase in performance	Average increase in performance
Most efficient power saving	Less efficient power saving

Comparing actual effect of both technologies:

Intel i7 Processors with Turbo Boost	Clock Speed	AMD FX processors with Turbo Core	Clock Speed
i7 – 960	3.20 GHz up to 3.46 GHz	FX 8150	3.6/4.2 GHz
i7 – 950	3.06 GHz up to 3.33 GHz	FX 8120	3.1/4.0 GHz
i7 – 920	2.66 GHz up to 2.93 GHz	FX 6100	3.3/3.9 GHz

Conclusion:

Intel leads the way in power efficiency and the Intel Turbo Boost is better in comparison to AMD in power saving and performance increase.

III. Pipelining of Intel Vs. Pipelining of AMD

Pipeline of Intel	Pipeline of AMD
One single efficient pipeline	3 Parallel pipelines
Single pipeline for all data structures	Different pipeline for different data structures
Macro and Micro fusion technologies	No such technologies
Loop stream detector block	No such block
2nd level branch predictor	Only one branch predictor
Stores history of branch prediction	No history storage

Conclusion: Intel has very sophisticated pipelining than AMD.

IV. Intel QPI Vs. AMD HyperTransport

Intel QPI	AMD HyperTransport
Link width is 20 bits	Link width is 32 bits
Unidirectional Send and Receive	Bidirectional Send and Receive
Non -Switchable link width	Switchable link width from 2 to 32 bit
No power management feature	Also facilitates power management

Comparison of both technologies:

Frequency	Intel QPI (rate)
2.40 GHz	19.2 GB/s
2.93 GHz	23.44 GB/s
3.20 GHz	25.6 GB/s

Frequency	AMD HyperTransport(rate)
1.40 GHz	22.4 GB/s
2.60 GHz	41.6 GB/s
3.20 GHz	51.2 GB/s

Conclusion:

From the above comparison we see that the AMD HyperTransport technology has exploited more in the processors and I/O interconnect as compared to Intel QPI and is better than the later one.

V. Intel Smart Cache Vs. AMD Balanced Smart Cache

Intel Smart Cache	AMD Balanced Smart Cache
LLC Shared across all cores, GPU and other units	LLC Shared across cores only
LLC is inclusive	LLC is non
Micro-op cache for decoded micro-op	None

Comparison of both technologies:

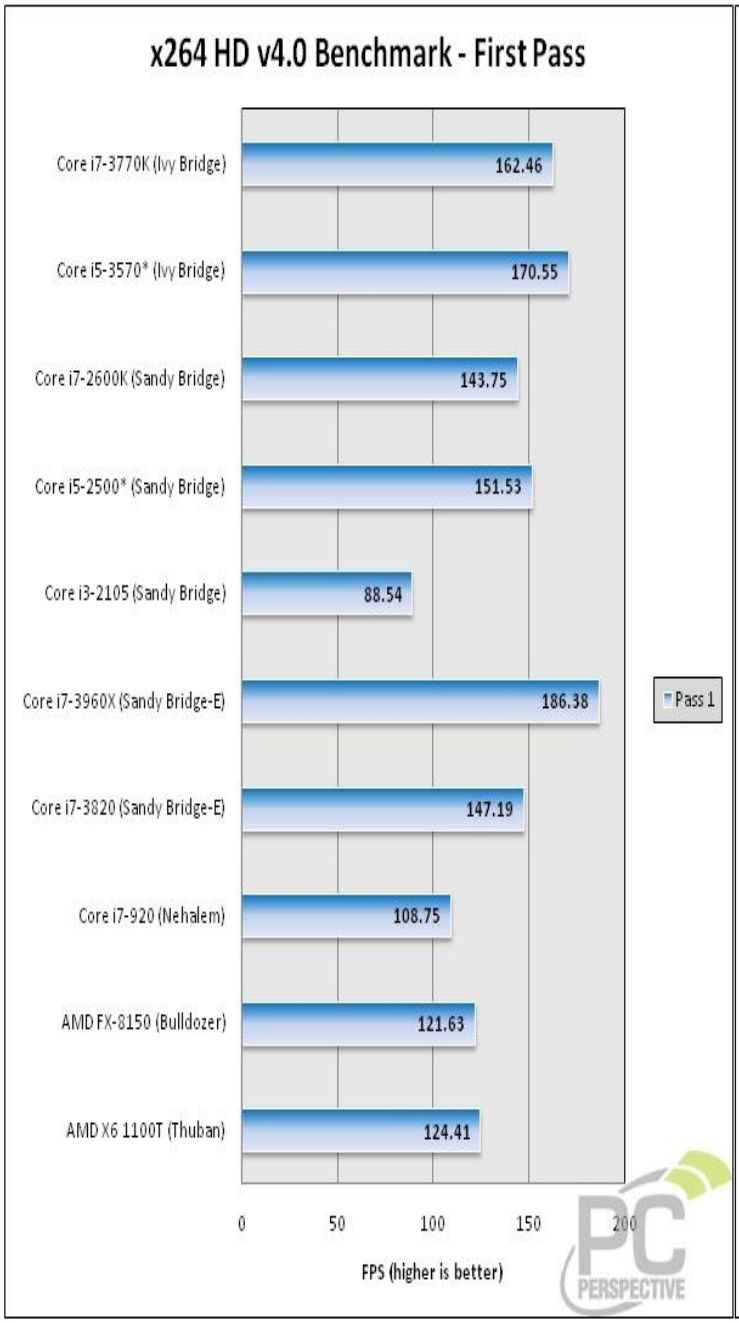
Cache	Core 2	Nehalem	Sandy Bridge	AMD(Bobcat)
Level 1 code	latency 3	latency 4	latency 4	latency 3
Level 1 data	latency 3	latency 4	latency 4	Latency 3
Level 2	latency 15	latency 11	latency 12	Latency 17-20
Level 3	-	latency 38	latency 20	-

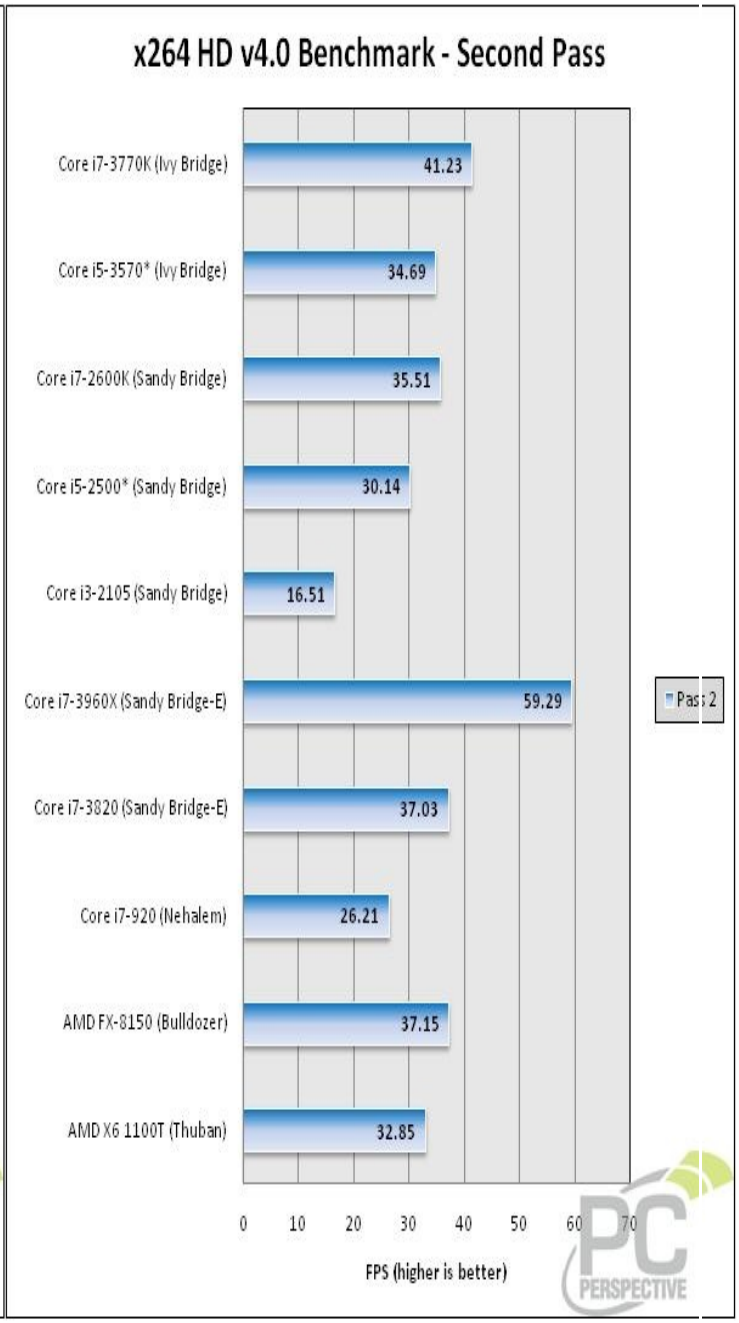
Conclusion:

Cache technology used by Intel is better than AMD as it doesn't have snooping time and also for fat execution of micro-op using Micro-op cache.

VI. Media Encoding Performance:

The popular x264 benchmark available from graysky does a two-pass H.264 encode on 720p video. The first pass appears to be more clock sensitive while the second pass is more heavily threaded.



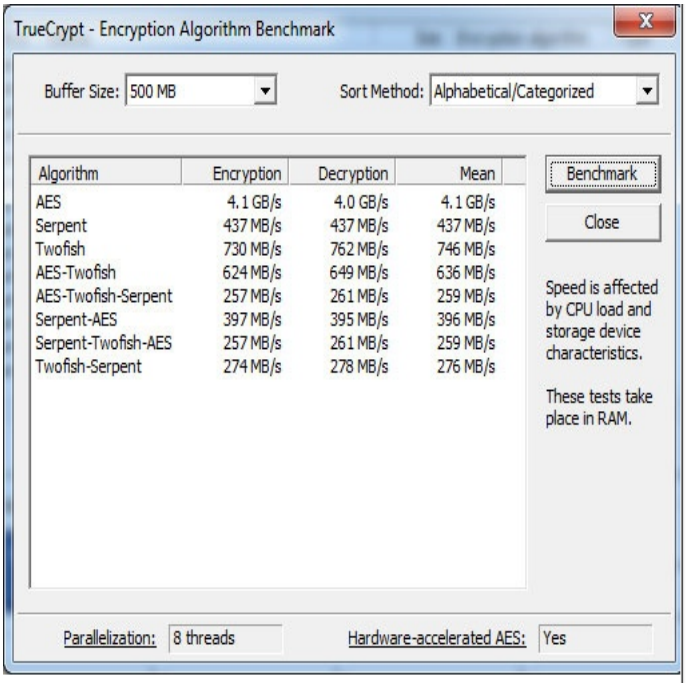


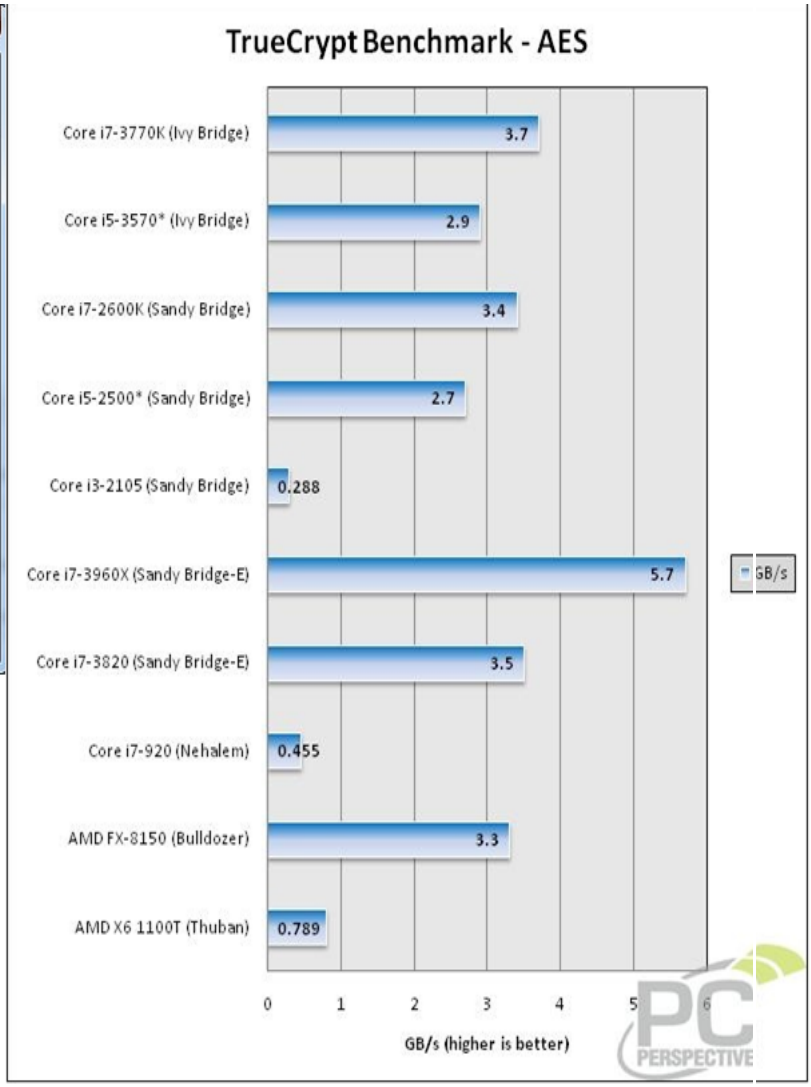
Comparison result:

1. The first pass shows the same kind of results that Handbrake did - higher performance with Hyper Threading disabled.
2. The second pass is very different though as we see nearly a 20% gain in performance by enabling it.
3. The i7-3770K is 16% faster than the i7-2600K and 11% faster than the Core i7-3820 SNB-E.

VII. Compression and AES Instruction Performance:

TrueCrypt is one of the most popular pieces of software for disk encryption and it includes a handy benchmark mode to test the capabilities of your processor. Keep in mind that many modern CPUs include AES acceleration that is used throughout TrueCrypt.

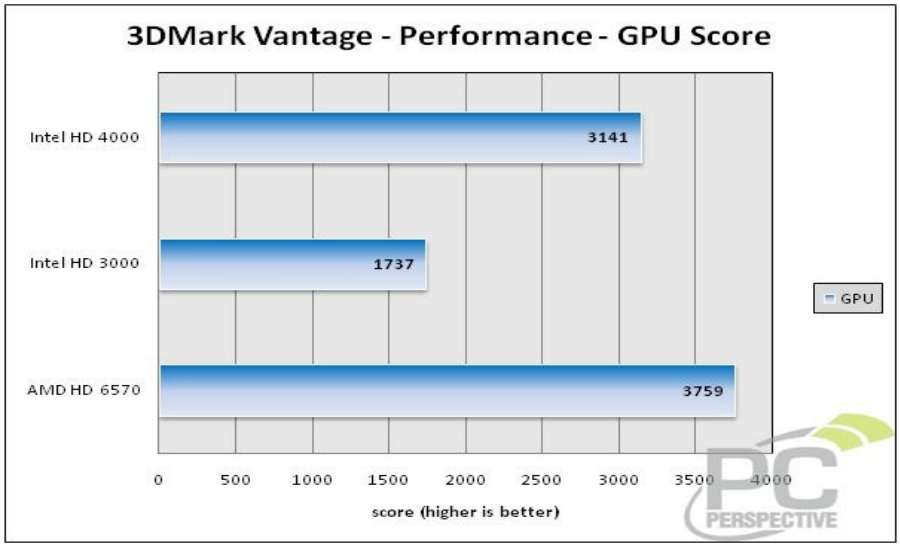




Comparison results:

1. The new Core i7-3770K is able handle 3.7 GB/s of encryption and decryption falling behind only the Core i7-3960X at 5.7 GB/s.
2. The new Ivy Bridge is able maintain a 19% performance lead over Sandy Bridge.

VIII. HD Graphics and QuickSync Performance:

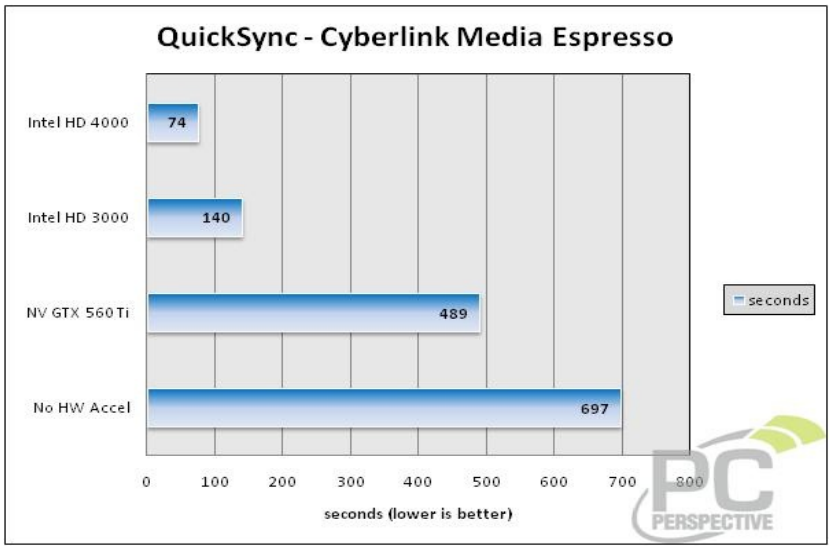


Comparison result:

1. Result here is the new i7-3770K is able to outperform the Core i7-2600K by more than 80% and stay within 19% of the discrete HD 6570 card.
2. Without a doubt the new Intel HD 4000 graphics are a big step up in performance for integrated graphics gaming. Seeing nearly 50% gains in L4D2 and DiRT 3 really show us that the Ivy Bridge processors could be competent enough to be the sole gaming solution for the casual gamer.

Quick Sync Performance:

For our basic Quick Sync testing we used the latest version of Cyberlink's MediaShow Espresso to convert a 1.8GB 720p H.264 video file down to a more reasonable size for an iPod.

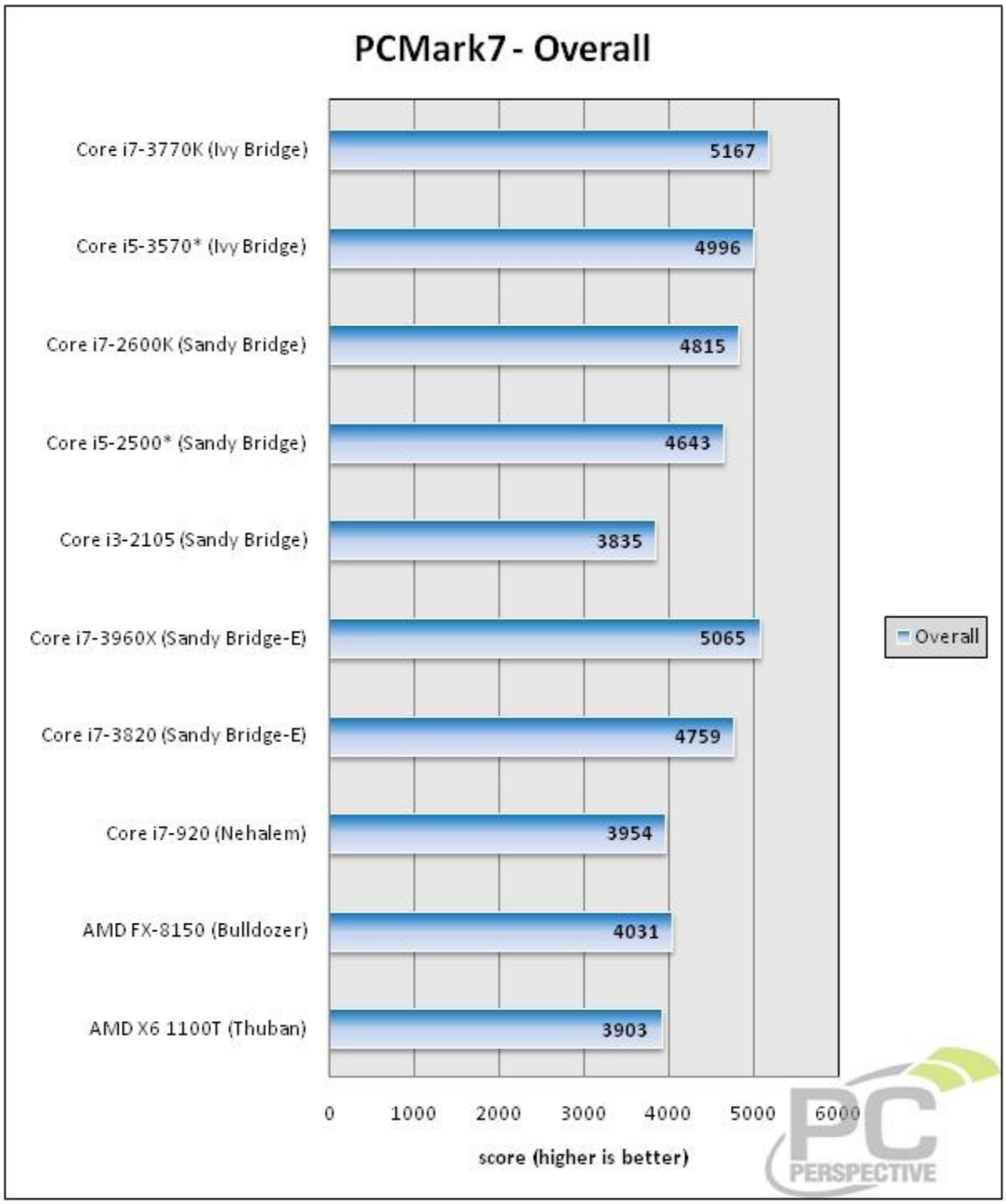


Comparison result:

Time required to convert was reduced to half in Ivy Bridge as compared to others.

C) Overall Performance:

PCMark 7 includes 7 PC tests for Windows 7, combining more than 25 individual workloads covering storage, computation, image and video manipulation, web browsing and gaming. Specifically designed to cover the full range of PC hardware from netbooks and tablets to notebooks and desktops, PCMark 7 offers complete PC performance testing for Windows 7 for home and business use.



Comparison result:

1. New 3770K Ivy Bridge beating out the SNB-E 3960X.
2. With some emphasis on the lightweight and productivity scores, the new Ivy Bridge processors look to do quite well not only for our compute-intensive benchmarks but also the standard computing workloads we all have.

5. CONCLUSION:

a. The report covers almost all the major technological advancements in the field of processors. No doubt the field is itself vast and requires an intense study. However the scope for technological improvements is ever high.

b. The third generation core i processors for desktop computing with “Ivy Bridge” microarchitecture defeats all its predecessors and also its competitor AMD processors.

c. To substantiate this, a 100 core processor chip has been fabricated by Prof. Anant Agarwal (MIT) for the servers. That is a quantum leap for make more efficient servers with better power saving and performance. The 100 cores will be accomplished by assembling multiple cores into one single chip.

d. The traditional "connecting bus" is replaced by a "switch" in new processor. According to Prof.Anant, "every processor has a switch and they all talk to each other like in a peer to peer network."

e. The new chip technology with 64 cores has been tested on the Facebook by its engineers. It yielded 67% higher throughput as compared to the Intel's Xeon and AMD's Opteron chips. The new 100 core chips are slated to be shipped later this year, promise an entire new era of computing.

f. The future of the computing world is bright, where more and more performance boost up technologies will be tested to their limit, till an entirely new technology is evolved.

6. References:

- http://www.intel.com/content/www/us/en/contentlibrary.html
- http://developer.intel.com/technology/itj/
- http://www.amd.com/in/Pages/AMDHomePage.aspx
- http://developer.amd.com/pages/default.aspx
- http://www.xbitlabs.com
- http://realworldtech.com/
- http://www.pcper.com
- http://www.extremetech.com
- http://en.wikipedia.org/wiki