AMD Fusion APU: Llano

Marcello Dionisio, Roman Fedorov

Advanced Computer Architectures
Outline

• Introduction
• AMD Llano architecture
• AMD Llano CPU core
• AMD Llano GPU
• Memory access management
• Turbo core technology
• Advanced power management
• Conclusions and future work
Introduction

The continuous refinement in the integrated circuits production processes lead to the possibility to integrate in the same die area of an old CPU a lot of additioonal functionalities like, in particular, an entire graphic processing unit (GPU) and at the same time to reduce the power consumption w.r.t. a discrete peripherals solution. This new extended version of CPU is called Accelerated Processing Unit.

AMD is been one of the first big CPU manufacturers to produce APUs for the mass market. In 2006 its APU line was announced with the name of AMD Fusion. The first commercial entered in commerce in 2011: their codenames where Brazos and Llano. The Llano architecture is the current AMD’s solution for its whole desktop and mobile products.
The main guidelines in the design of the Llano architecture were the balancing between performance and power, optimization of performance for different workload classes and increasing of battery life and power savings.
The Llano architecture is composed by 3 main modules:

- **CPU**: 4 cpu x86 cores for general purpose activities
- **GPU**: an integrated DirectX 11 Graphic core
- **I/O, memory and display controllers**

All in a 227mm$^2$ area!

In the successive slides we’ll see each component in detail
AMD Llano CPU core (1)

Llano contains 4 x86 CPU cores derived from the AMD Stars architecture with the following improvements:

- Increasing of L2 cache memory up to 1 MegaByte per core
- A reording buffer capable of 84 micro-operations (better ILP)
- Extended load/store buffer and a more aggressive load/store forwarding policy (better MLP)
- Enhanced instruction prefetcher (better IPC)
- Enhanced floating point scheduler

All the performances related to the CPU components are digitally monitored in order to balance the power consumption.
The Llano core implements the power gating ring, that allows to power each core individually, so optimizing the power savings with an arbitrary number of active threads.
The Llano architecture includes a DirectX 11 graphic processing unit which uses the third generation of the Unified Video Decoder (UVD-3) for 3D and 2D video playback.

Llano integrates 8 PCI-express 16x lanes for an external graphic card, 8 lanes for the display port of the integrated card and 16 lanes which can be used by both the external graphic card and the display port.
The integrated GPU includes all the features implemented in AMD discrete graphic cards with low-latency, high bandwidth and access to the shared system memory:

- 16 AMD Radeon VLIW-5 cores composed by 4 stream cores, a special function core, a branch unit and some general purpose registers to co-issue multiple operation in a single clock cycle.
- Each of these sets of VLIW cores forms a SIMD (single instruction multiple data) unit. The Llano GPU core contains 5 SIMD units capable of a throughput of 480 Gflops.
The power controller manages the power consumption of the GPU with no external driver intervention. It completely shuts down the graphic core when it seems to be idle for a time interval longer than a programmed one and, when it’s needed again it restores the power and the context of the GPU. All these operation are performed in less than 10 ms.
Llano implements UMA (Unified Memory Architecture): processor and graphics share a common memory. A portion of the memory can be also dedicated to graphics buffer memory.

All traffic of graphics nature is routed through the GMC (Graphic Memory Controller) that builds a stream of memory requests to the NorthBridge.

The display controller keeps a local buffer with the next few lines to be sent to the panel, and to avoid the buffer underrun (that will cause the tearing effect) the DRAM controller prioritizes the display effects over other requests when the buffer is almost empty.
Other graphics related streams can be handled in GMC or DRAM controller, which ensures that the average latency of requests never goes above a configured threshold for each requests type.

Processor and I/O traffic streams are managed in the NorthBridge front-end stage, and gets allocated a proper bandwidth.

A mechanisms of the DRAM controller called Starvation timers check the latency of the stream and allow forward progress in case the requests are stalled longer that the defined (dynamically) threshold. In this way the DRAM controller can guarantee the minimum required bandwidth for each stream.
Turbo core technology (1)

It performs the tracking of power consumed by an application and boosts (or decreases) the frequency of the APU components in order to respect the TDP (thermal design power) limit. The TDP can dynamically be allocated to different APU components (e.g. For CPU-centric or GPU-centric applications). ATC introduces the measurements of instantaneous energy margin and energy margin at time k (EM(k)) to balance the power distribution among the cores.

\[
\text{IEM} = \text{TDP Limit} - \sum_{i=0}^{4} (\text{Compute Unit Power}(i))
\]

where

\[
\text{Compute Unit Power}(i) = P_{\text{Leakage}} + C_{ac} V^2 f
\]

\[
\text{EM}(k) = \sum_{i=\text{INT}(kt/T) \times T/t}^{k} \text{IEM}(i) + D \times \text{EM}(\text{INT}(kt/T) \times T/t)
\]
Another concept introduced by ATC is “variable core TDP limit”:

The ability of a core to dissipate heat depends on what is happening to the adjacent cores activity:

- A single core can benefit a significantly high TDP while using the other cores as heat sinks
- Other combinations are allowed (dual core tasks, GPU centric tasks…)

![Diagram showing variable core TDP limit concepts](image-url)
Advanced Power Management (1)

Different sets of policies of Llano PMC:

- The max battery/DC source policies:
  Maximize battery life
- The max performance/AC source:
  Prevent any performance loss.

Apart C and P states defined by ACPI there are two low-power states – power gating (CC6) and clock off.

CC6 – long entry and exit times (due to the energy cost of flushing the cache and saving/restoring content), so not profitable to use during high system activity.
PCM makes dynamic decisions on the appropriate power state for each core based on a set of configurable monitors:

- Active/idle residency trackers
- DMA tracker
- Timer tick interval updates
- Interrupt-rate tracker (IRT)

In practice the IRT is the most efficient monitor and it is widely used by the PMC to take power state decisions
Conclusion and future work

Main challenges in the future work:

- Enrich the amount of functionalities integrated in the APU (I/O engines and controller)
- Improve the performance/watt ratio depending on different workloads
  - Better tracking of application type
  - Better performance sensitivity
- Integrate also off-die (e.g. discrete GPU) entities in the application management.