IMEC Middleware:

Maximizing Throughput on Barco’s GPU-Enabled Video Processing Server

A Step towards Vision in the Cloud

Maja d’Hondt

IMEC
Overview

• Imec – NVision – Ares
• Professional video processing on commodity server – a project with Barco
  ▸ Challenges
  ▸ Imec middleware
  ▸ Results
• Next step: multiple servers
• The future: Vision in the Cloud
IMEC 1984 – 2010

1984

- Established by state government of Flanders in Belgium
- Non-profit organization
- Initial investment: 62M€
- Initial staff: ~70

2010

- World-leading research in nanoelectronics
- Revenue: 275 M€ (incl. 44 M€ grant from Flanders government)
- Staff: > 1750 worldwide
- Worldwide collaboration >600 companies
- Research 3-8 years before product
MISSION

Imec performs world-leading research in **nanoelectronics**.

We deliver **industry-relevant** technology **solutions**.

We leverage our scientific knowledge with the innovative power of our **global partnerships** in **ICT**, **healthcare** and **energy**.
**IMEC Business Lines**

<table>
<thead>
<tr>
<th>IMEC Core CMOS</th>
<th>IMEC CMore</th>
<th>Human++</th>
<th>IMEC Energy</th>
<th>IMEC Smart Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lithography</td>
<td>SiGe MEMS</td>
<td>Wearable and implantable body area networks (with Holst Centre)</td>
<td>Photovoltaics</td>
<td>Power-efficient green radios</td>
</tr>
<tr>
<td>Logic DRAM devices</td>
<td>Silicon photonics</td>
<td>Vision systems</td>
<td>GaN power electronics and LEDs</td>
<td>Large-area electronics and systems-in-foil (with Holst Centre)</td>
</tr>
<tr>
<td>Interconnects</td>
<td></td>
<td>Power devices and mixed-signal technologies</td>
<td></td>
<td>Vision systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GaN power electronics and LEDs</td>
<td></td>
<td>Wireless autonomous transducer solutions (with Holst Centre)</td>
</tr>
</tbody>
</table>

Emerging devices
INSITE – connecting technology and system design
Life sciences
NVision – Past

Video processing and 3D applications
Compilers, tools, middleware
Platforms and processors
NVision – Next-generation Vision Systems

- Smart Lenses
- Volumetric Displays
- Compact Vision Systems
- ARES: Middleware for professional video processing on heterogeneous commodity hardware
- Adaptive Optics
- Hyperspectral Imaging
Broadcasting on Commodity Hardware

- professional display generators for video walls
- from dedicated DSP-based devices to commodity hardware (CPU and GPU)
  - inside the box: CPU and GPU
  - outside the box: networked commodity servers
Focus: inside the box

• why move to commodity processors?

› dedicated processors:
  – fixed function – no flexibility
  – overdimensioning
  – bottlenecks and idle

› processing is very dynamic
  – different video stream quality
  – different number of video streams
  – different processing, e.g. depending on video analysis

› flexibility and scalability
  – load balancing
  – increase throughput
# Challenges for Ares Middleware

- No more fixed function components

## Challenges vs. Ares Middleware

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Ares Middleware</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterogeneous processors and ≠ data transfer times</td>
<td>Integrated model for load balancing (monitoring and migration)</td>
</tr>
<tr>
<td>No additional design time</td>
<td>Run-time monitoring</td>
</tr>
<tr>
<td>Variable workloads</td>
<td>Run-time migration</td>
</tr>
<tr>
<td>Portability</td>
<td></td>
</tr>
<tr>
<td>Low latency and no visual artefacts</td>
<td>Smart load balancing strategies</td>
</tr>
<tr>
<td></td>
<td>Optimized migration</td>
</tr>
<tr>
<td></td>
<td>Negligible overhead of 0.05%</td>
</tr>
</tbody>
</table>

30% more throughput wrt fixed function strategy on Barco’s video processing servers
VIDEO PROCESSING PIPELINES

- pipeline consists of components
- encoders, decoders, transcoders, scalers, analysis, …
- 3D components

Diagram:
- CPU:
  - src → encode → motion estimation → file

- GPU:
  - src → decode → analyze

  - transcode → screen → file
Pipelines in GStreamer

- GStreamer plugin contains both x86 and CUDA versions
- compatible with OpenCL, compilers, tools that automatically generate from one code base to both x86 and CUDA
Ares Middleware

• for each frame, for each pipeline component

• middleware will assign processing to certain processing element, i.e. GPU or CPU, at run time

• based on information monitored at run time
  ▸ processing time of pipeline component for one frame on each supported processing element
  ▸ data transfer time from CPU to GPU, and in some cases, back

• also based on availability (e.g. first free, fastest free)
Execution and Data Transfer Times

- timing predictor – different approaches
Motion Estimation (ME): compute and memory intensive algorithm, highly parallel.
**EXPERIMENT: PERFORMANCE wrt FIXED ASSIGNMENTS**

Pipeline components have **different best fixed assignment** on either CPU or GPU depending on **actual workload** of all running pipelines.

E.g. 6 different best fixed assignments for 8 different workloads.

Ares middleware performs almost always **better than each different best fixed assignment** per workload.

---

**GPU bottleneck**

different fixed assignments on CPU and GPU per workload
Experiment: Portability wrt Fixed Assignments

Variations in configurations have different best fixed assignments for same workloads:

- Fixed assignment A is best
- Fixed assignment B is best

Ares middleware: exact same software stack adapts to configuration and achieves best performance all the time (horizontal lines)
EXPERIMENT: INCREASED THROUGHPUT INSIDE BARGO’S VIDEO PROCESSING SERVER

# processed frames

# streams

ideal (no frame drop)
imec middleware
fixed function (CPU and GPU)
fixed function (CPU only)
EXPERIMENT: PROCESSED STREAMS AT DIFFERENT QUALITY LEVELS

- imec middleware
- fixed function (CPU and GPU)
- fixed function (CPU only)

capacity (# streams)

framedrop limit
Next Step: Outside the Box

- load balancing between servers
THE FUTURE: VISION IN THE CLOUD

• from pipelines and components to applications and services
• video processing and 2D/3D (compositing)
• for different terminals
• for different bandwidths
• elasticity
• optimal use of hardware
• power efficiency
Conclusions

• professional video processing is moving from dedicated devices, to commodity hardware, to the cloud
  ▸ quality – low latency and no visual artefacts
  ▸ no fixed-function – flexibility and scalability
• Ares middleware manages server processing resources for variable video processing workloads at run time
  ▸ heterogeneous load balancing
  ▸ monitoring
  ▸ pluggable timing predictors and strategies
• 30% increased throughput, 0.05% overhead, platform variability
• future: Vision in the Cloud