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 investement: 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**.

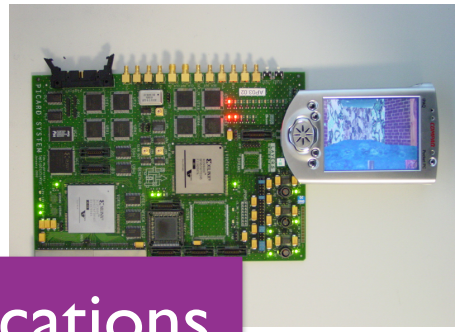
We deliver **industry-relevant** technology **solutions**.

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

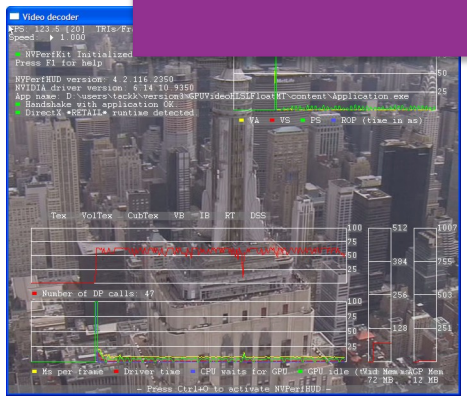
IMEC BUSINESS LINES

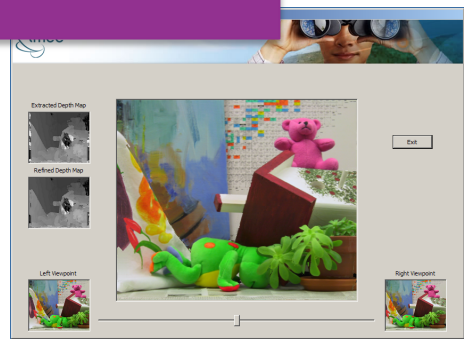
IMEC CORE CMOS Lithography Logic DRAM devices Interconnects	3D integration Flash memories	Emerging devices INSITE – connecting technology and system design
IMEC CMORE SiGe MEMS Silicon photonics	Vision systems Power devices and mixed-signal technologies	GaN power electronics and LEDs
HUMAN++ Wearable and implantable body area networks (with Holst Centre)	Life sciences	
IMEC ENERGY Photovoltaics	GaN power electronics and LEDs	
IMEC SMART SYSTEMS Power-efficient green radios Vision systems	Large-area electronics and systems-in-foil (with Holst Centre)	Wireless autonomous transducer solutions (with Holst Centre)

NVISION – PAST

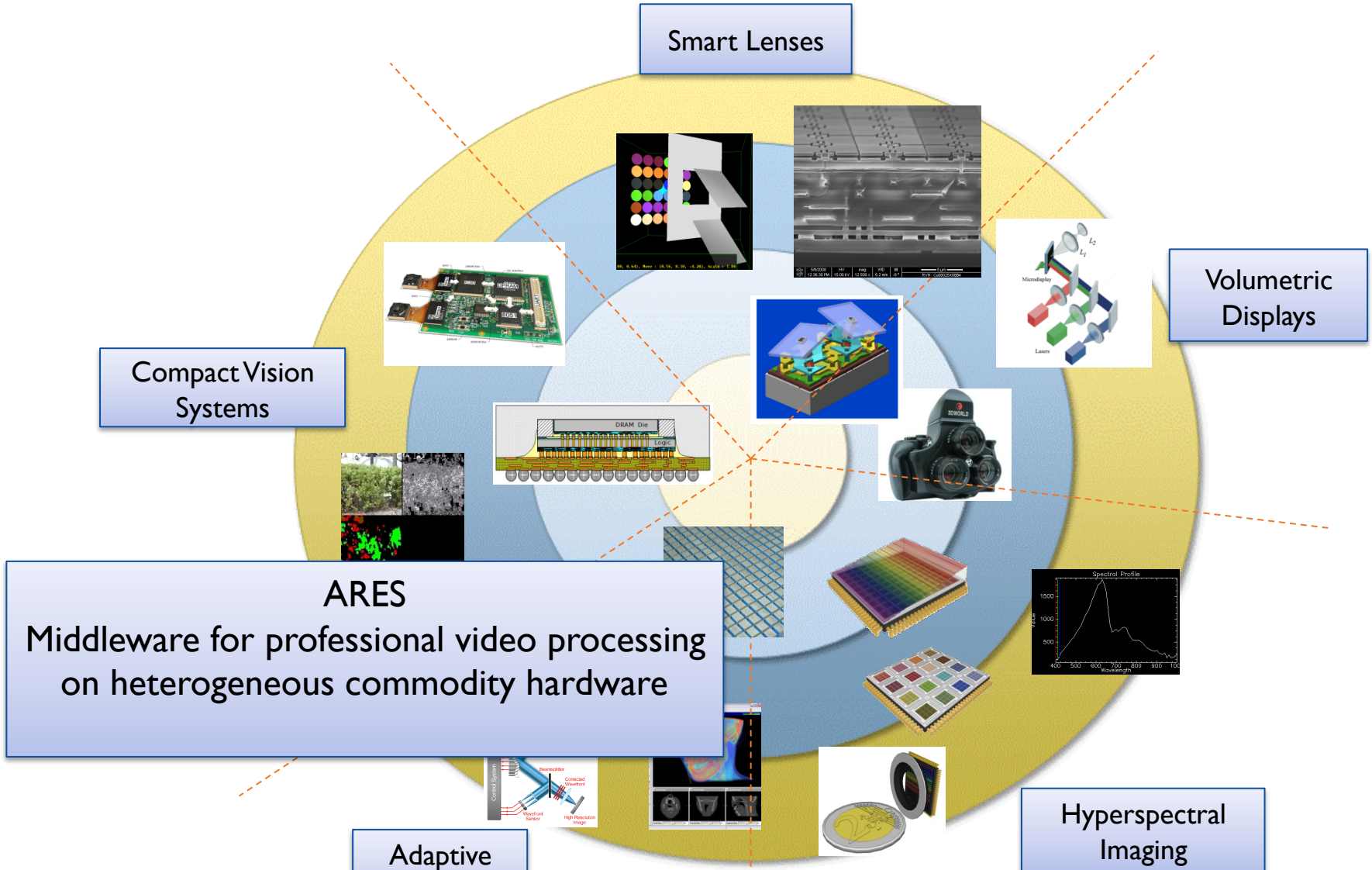


Video processing and 3D applications
Compilers, tools, middleware
Platforms and processors





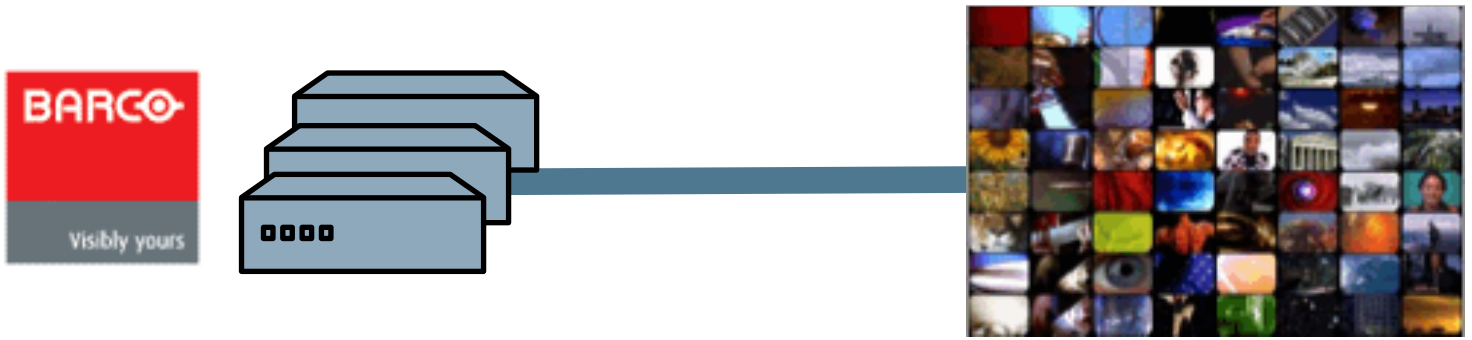
NVISION – NEXT-GENERATION VISION SYSTEMS



ARES
Middleware for professional video processing on heterogeneous commodity hardware

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 → load balancing

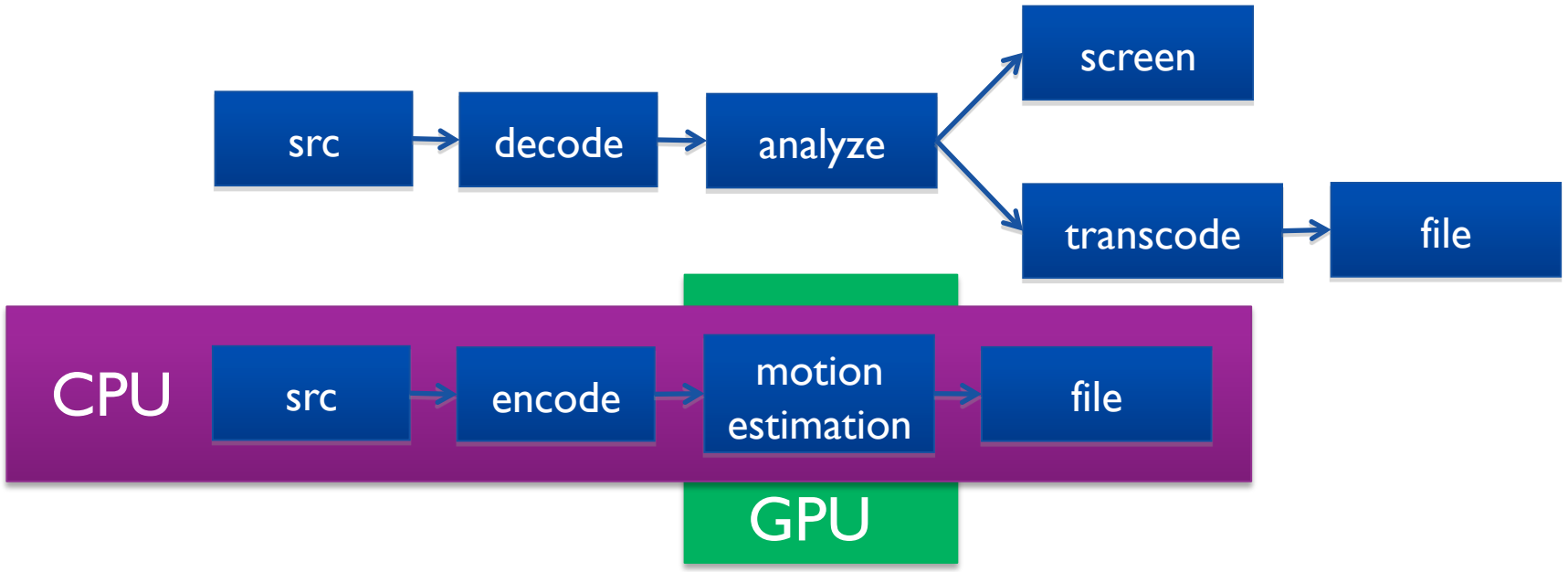
CHALLENGES	ARES MIDDLEWARE
heterogeneous processors and ≠ data transfer times	integrated model for load balancing (monitoring and migration)
no additional design time	run-time monitoring
variable workloads	
portability	run-time migration
low latency and no visual artefacts	smart load balancing strategies optimized migration neglibeable overhead of 0,05%

→ **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



PIPELINES IN GSTREAMER

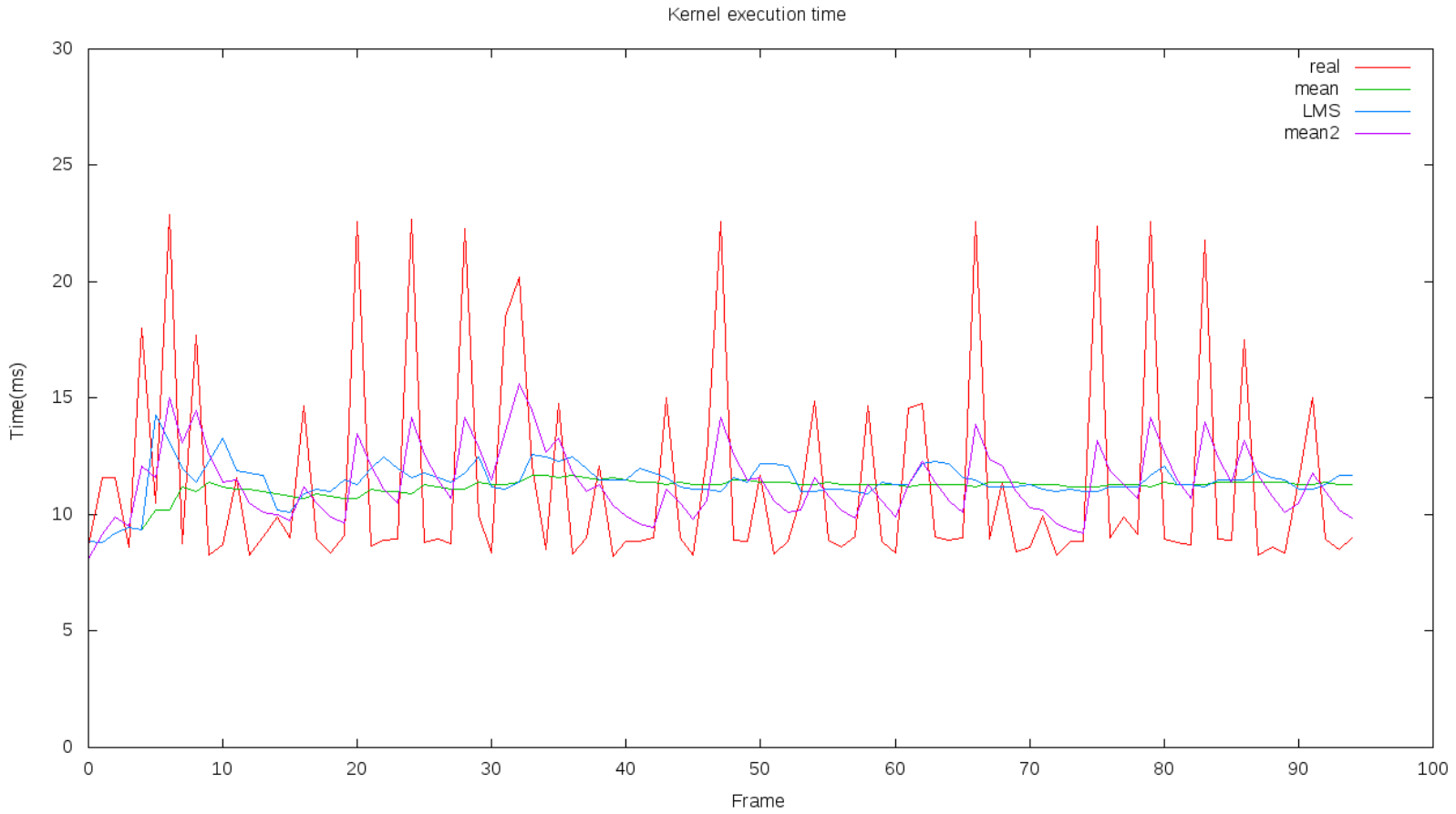
- GStreamer plugin contains both x86 and CUDA versions
- compatible with OpenCL, compilers, tools that automatically generate from one code base to both x86 and CUDA

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)

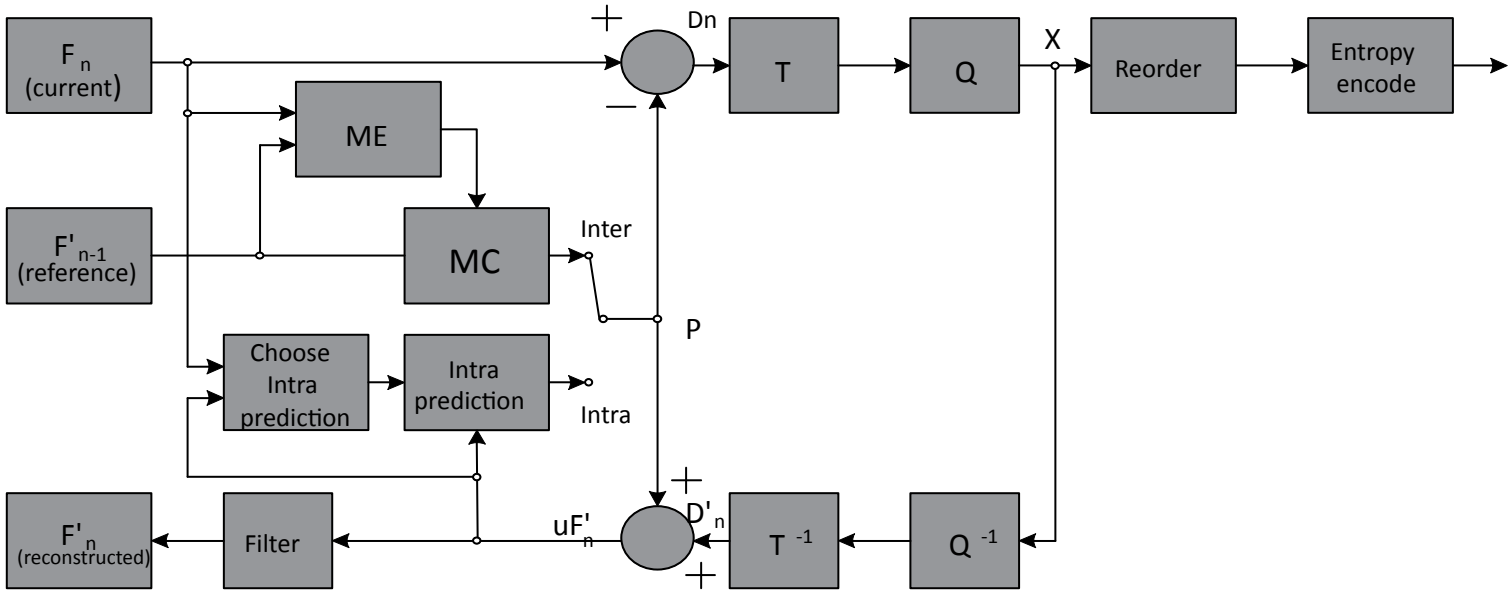
EXECUTION AND DATA TRANSFER TIMES

- timing predictor – different approaches



H.264 MOTION ESTIMATION IN CUDA

Motion Estimation (ME): compute and memory intensive algorithm, highly parallel.

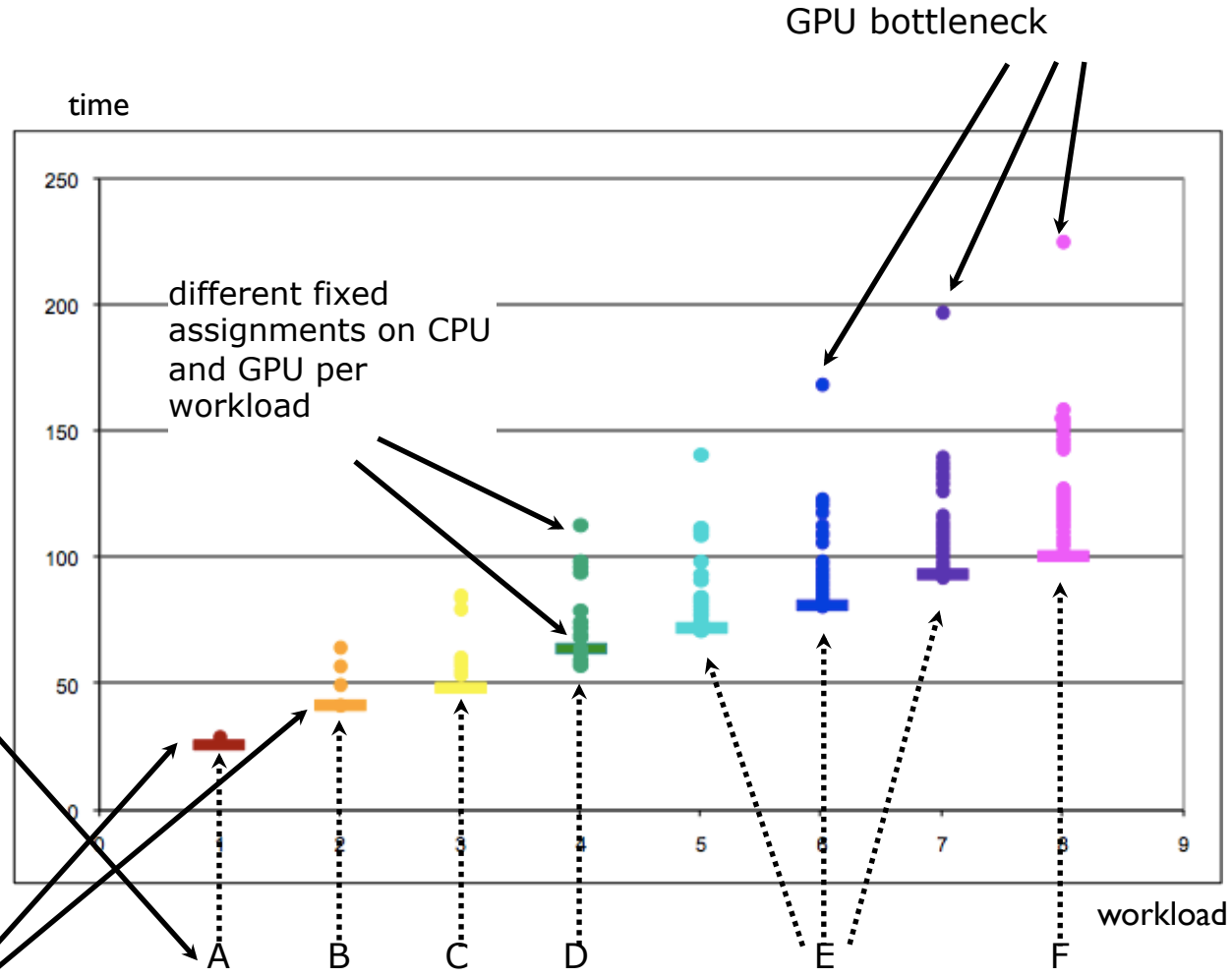


Experiment: Performance wrt Fixed Assignments

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

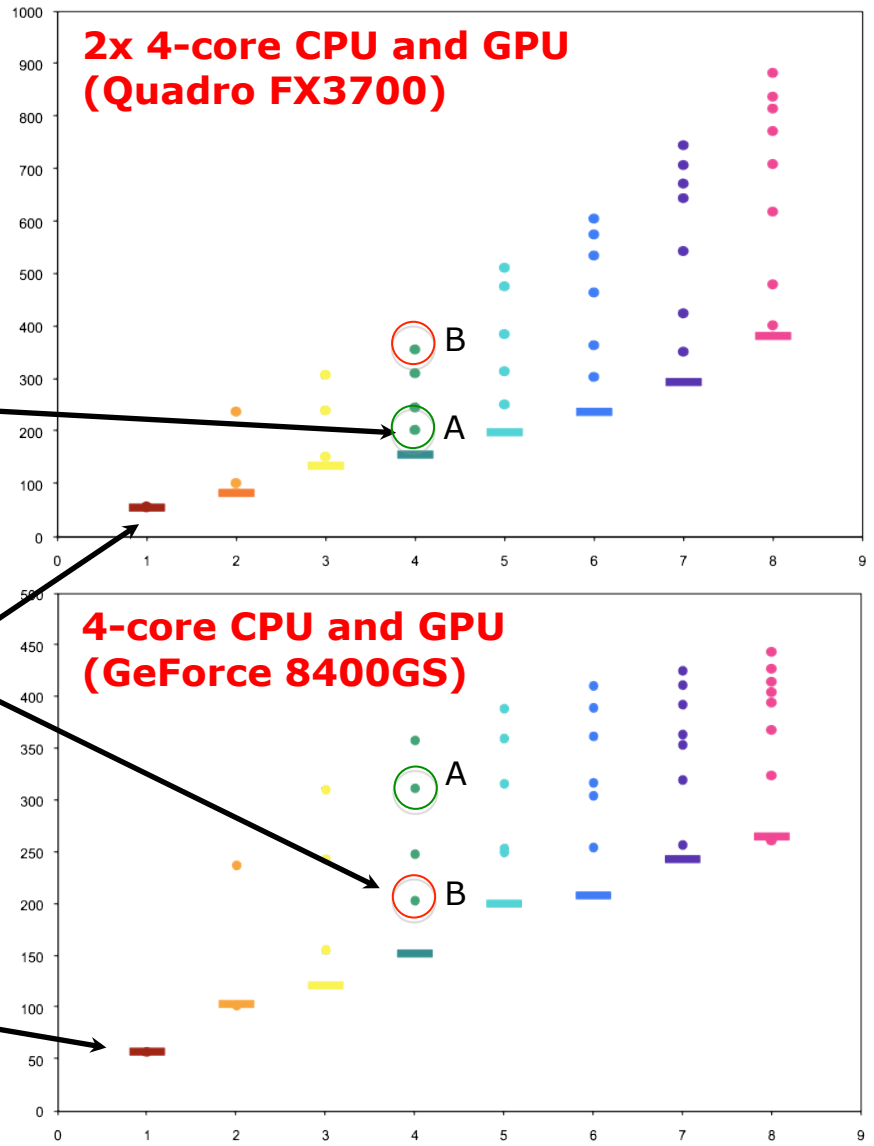
Ares middleware performs almost always **better than each different best fixed assignment** per workload



Experiment: Portability wrt Fixed Assignments

variations in configurations have different best fixed assignments for same workloads

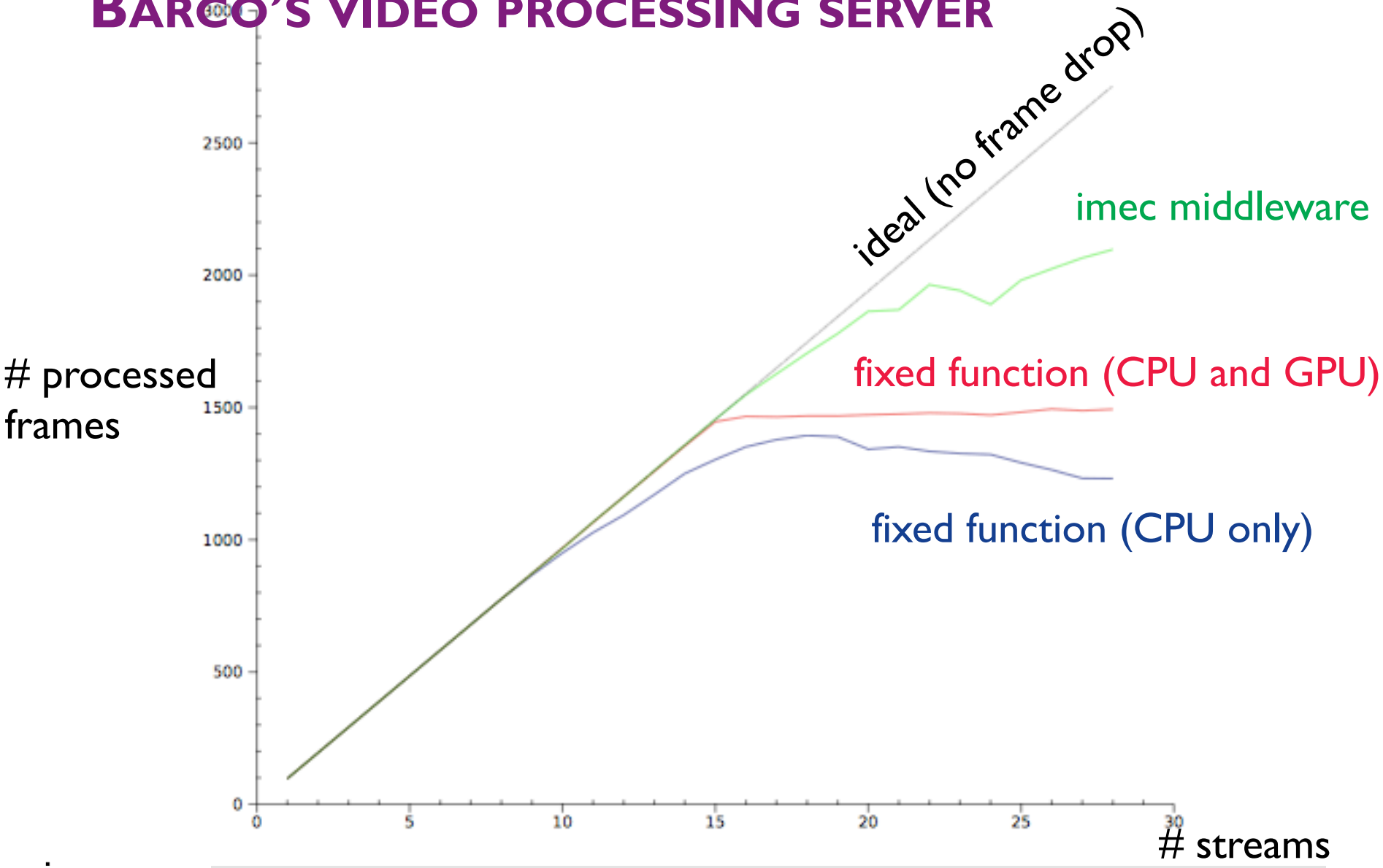
- fixed assignment A is best
- fixed assignment B is best



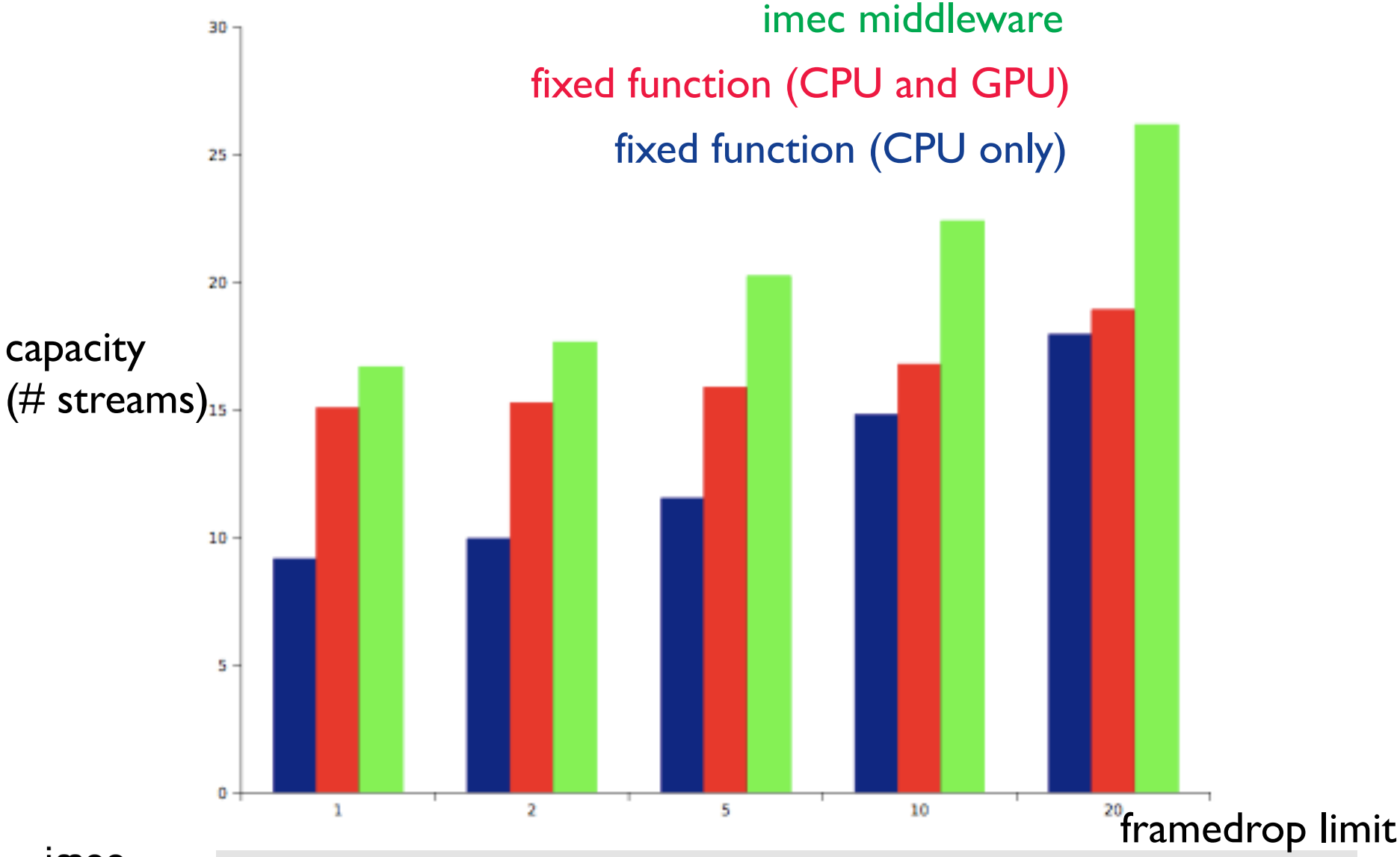
Ares middleware: exact same software stack **adapts** to configuration and achieves **best performance** all the time

(horizontal lines)

EXPERIMENT: INCREASED THROUGHPUT INSIDE BARCO'S VIDEO PROCESSING SERVER

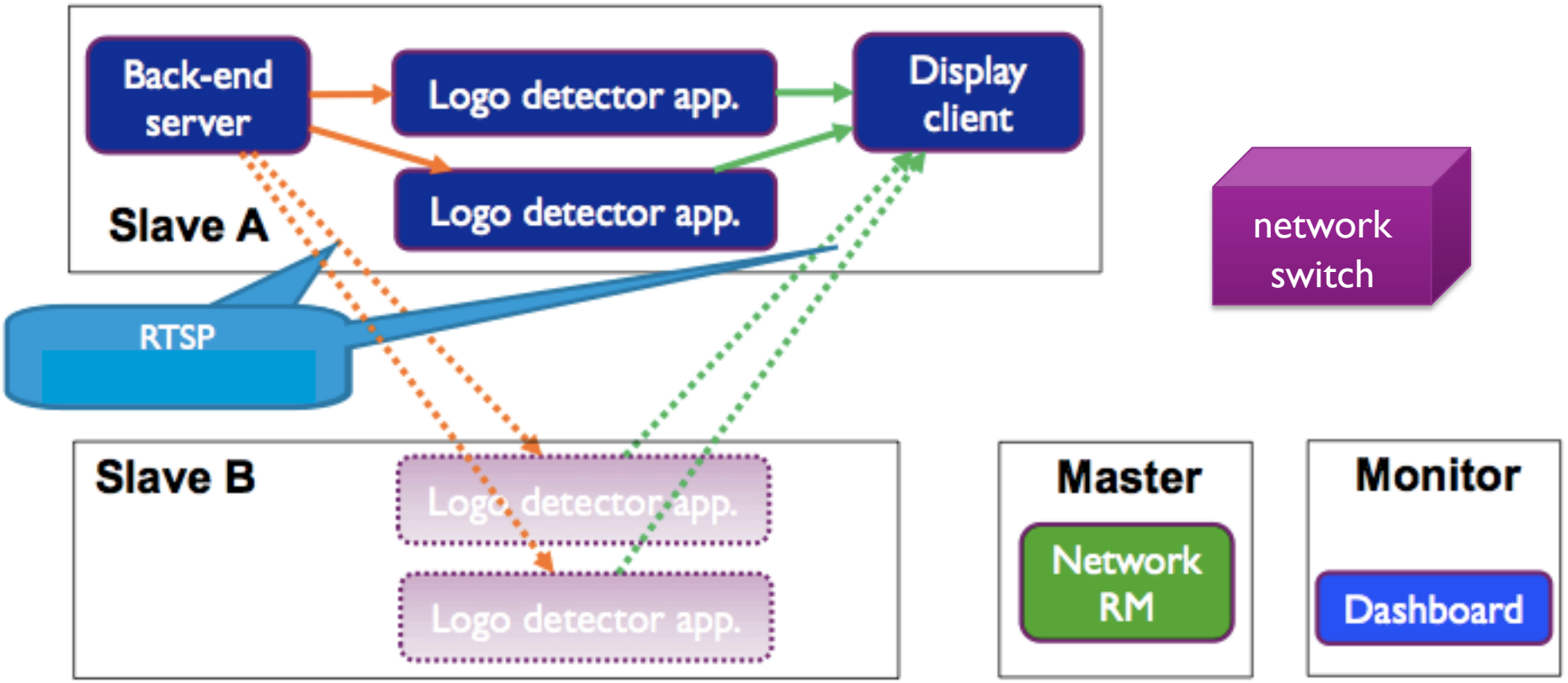


EXPERIMENT: PROCESSED STREAMS AT DIFFERENT QUALITY LEVELS



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