Low-Overhead Hard Real-time Aware Interconnect Network Router

Michel A. Kinsy

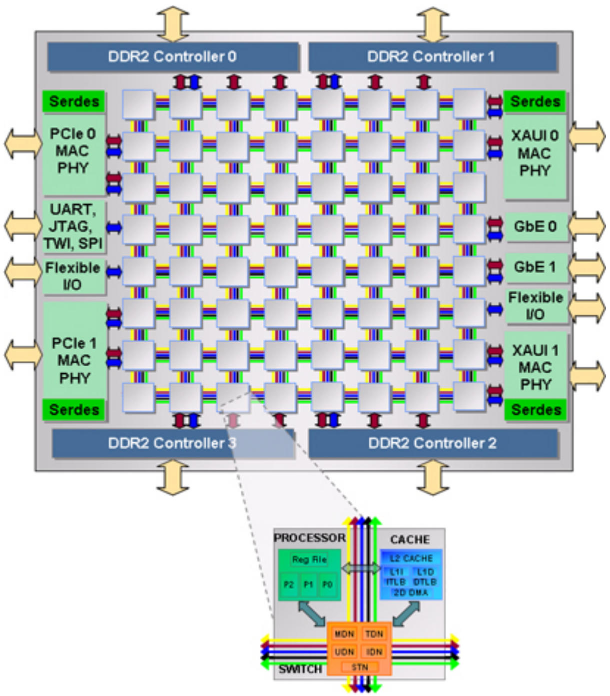
Department of Computer
and Information Science
University of Oregon

Srinivas Devadas

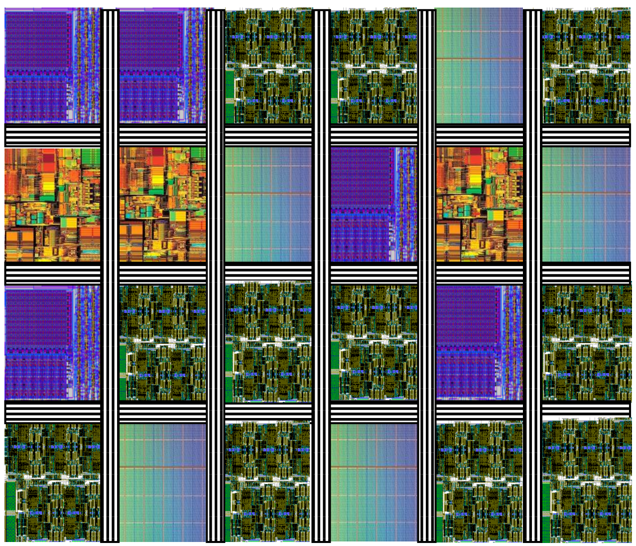
Department of Electrical Engineering
and Computer Science
Massachusetts Institute of Technology

Current System Trends

- **The convergence of application specific processors and multicore systems**
- **New phase where we see two distinctive computing architectures emerging**



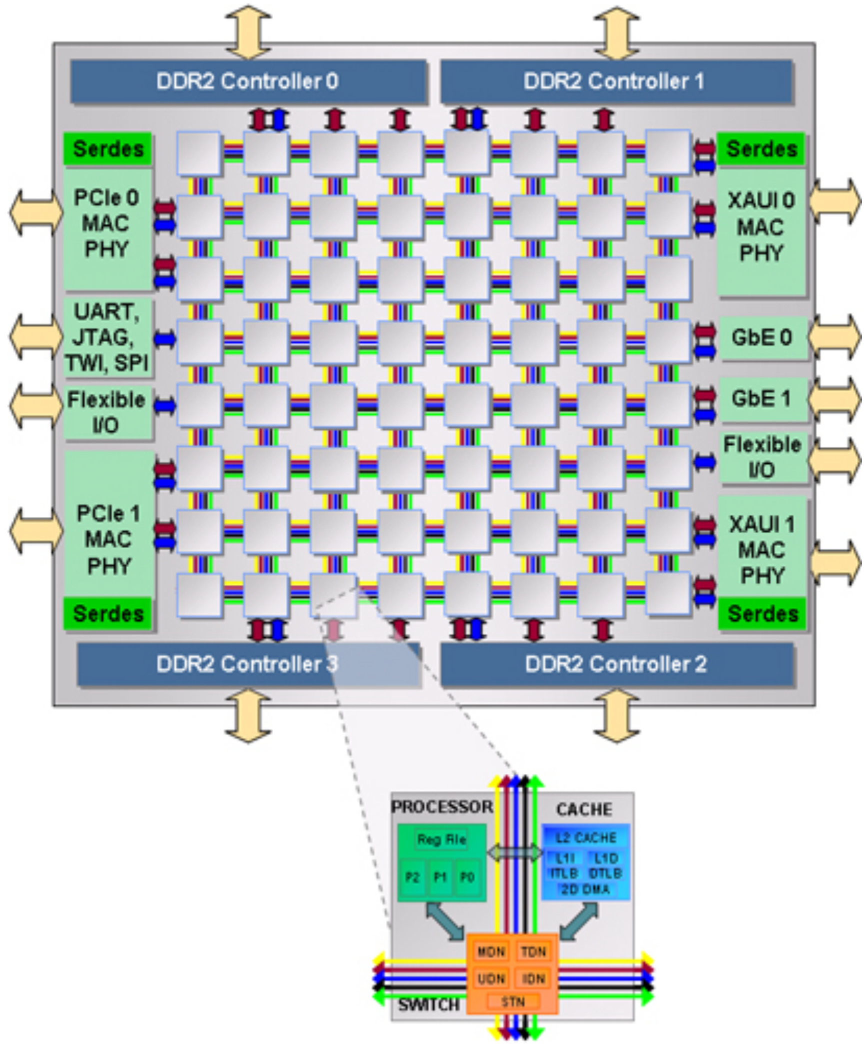
Massive parallel homogenous cores



Heterogeneous many-core architectures

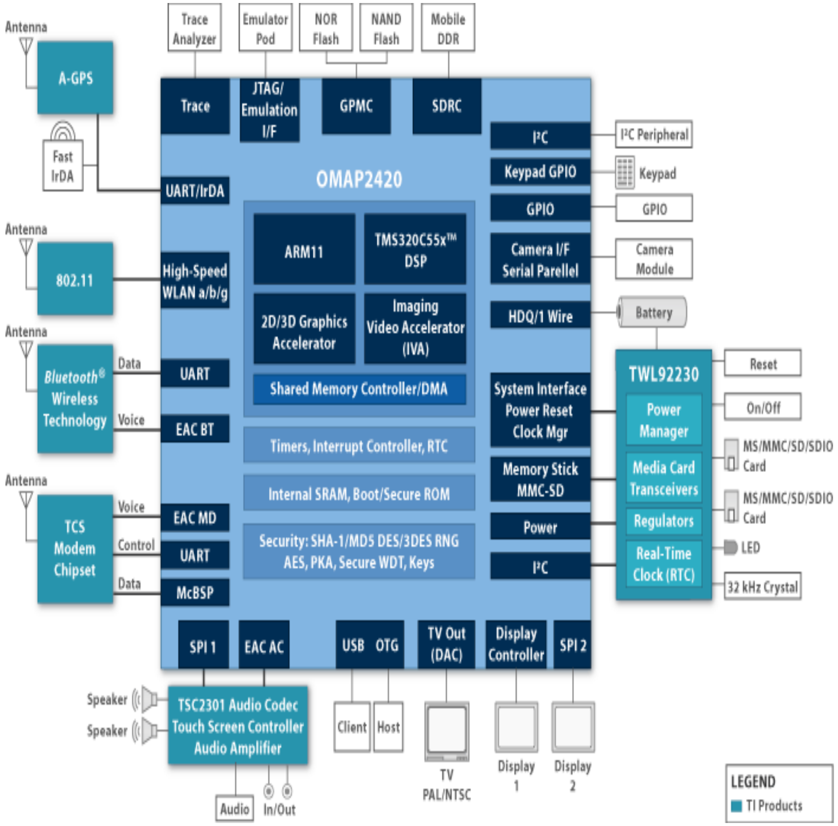
Homogeneous Many-cores

- **Tilera-like homogenous architectures: servers, cloud computing…**
- **Great for trivially parallel applications**
- **Having similar cores on the die makes the manufacturing and testing process more manageable**
- **Homogeneous collection of cores also keeps the software support model simple**
- **Lack of specialization of hardware to different tasks**



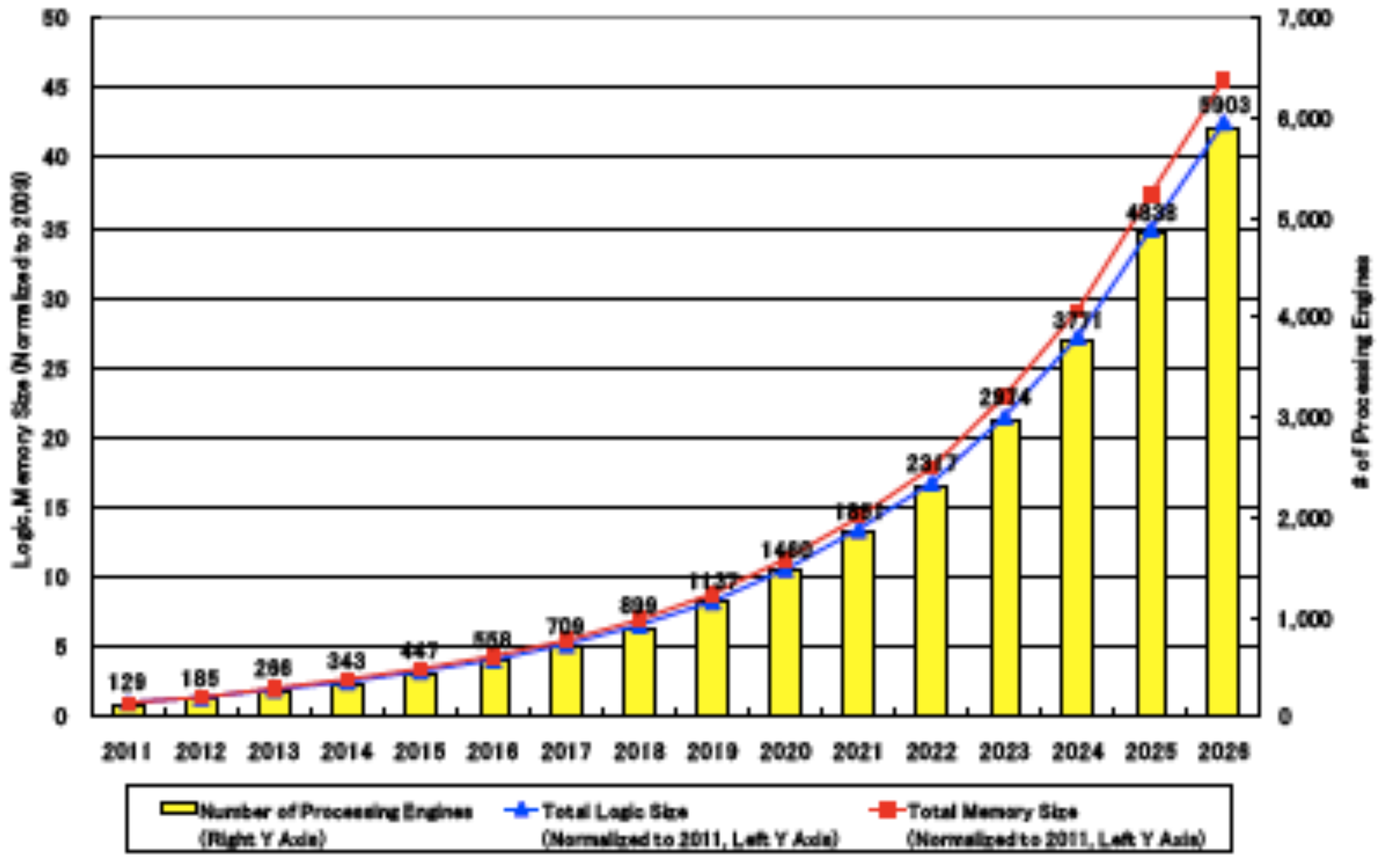
Heterogeneous Many-cores

- Integration of heterogeneous technologies
- These SoC architectures have large number of processing units
  – programmable RISC/CISC cores, memory, DSPs, and accelerator function units/ASIC
- Logic diagram of the hardware processing units in a typical smart phone today



Complex SoC Architectures

- SoC architectures need more research and standardization to encourage high degree of reusability



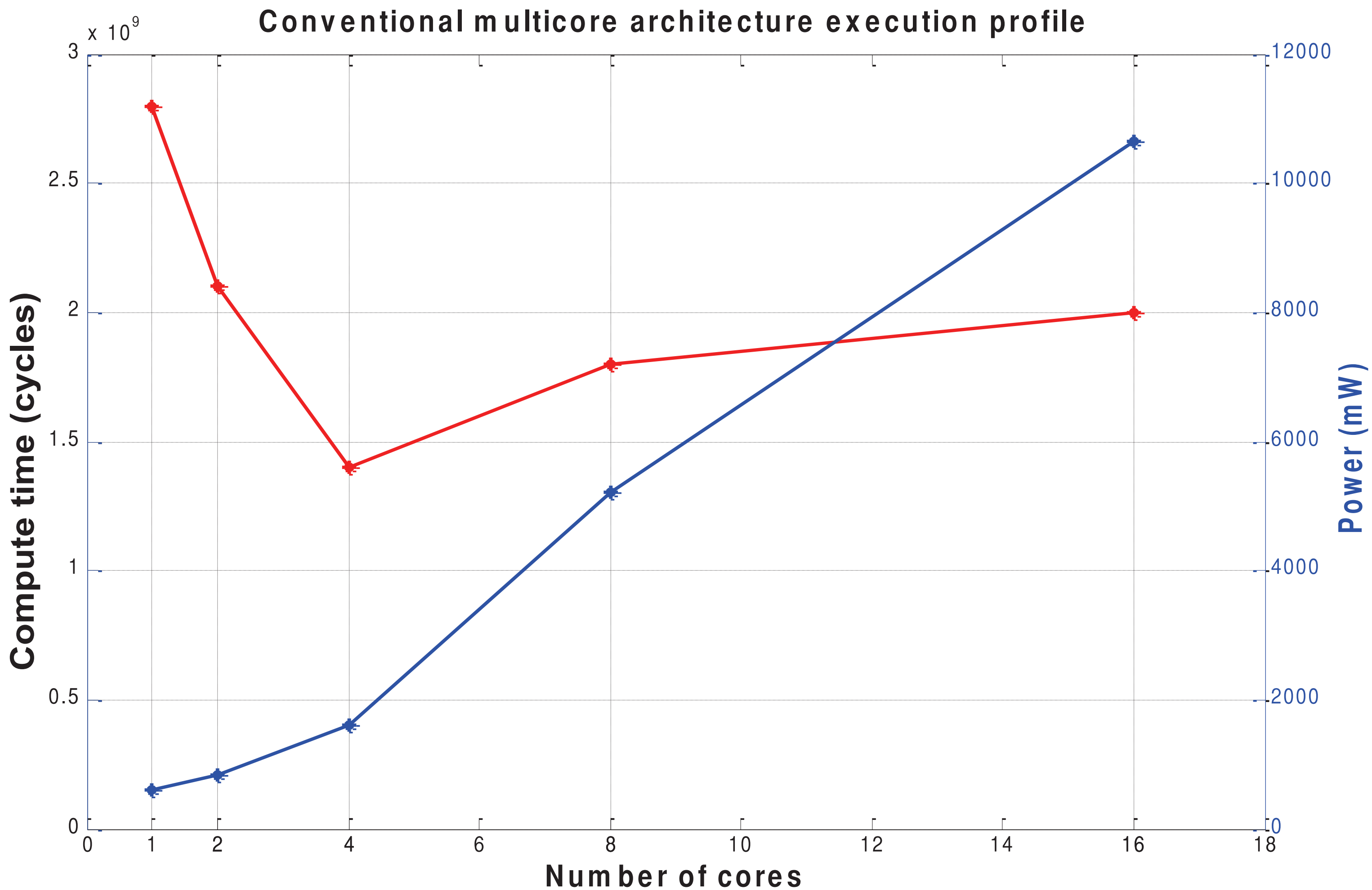
SoC design complexity trends [International Technology Roadmap for Semiconductors 2011 Report]

Design Challenges

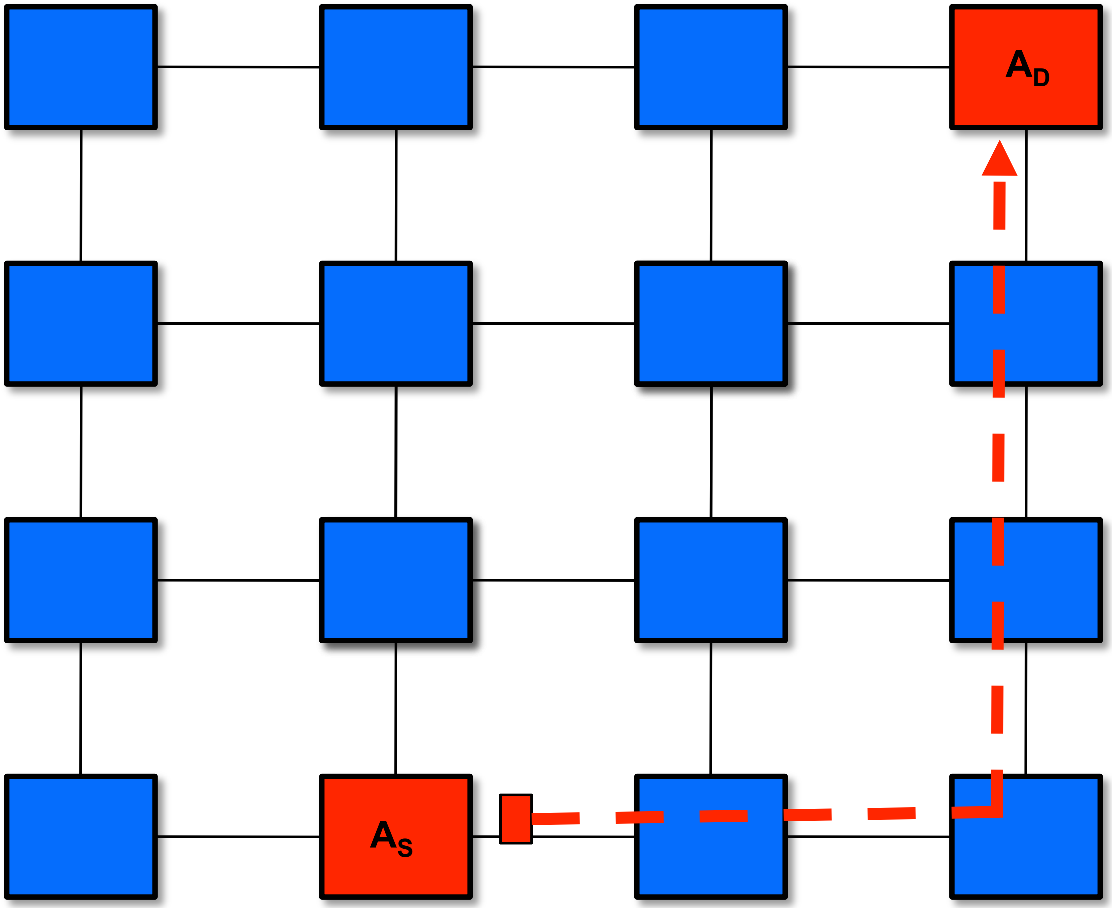
- Design of heterogeneous many-core systems shares many of the same challenges in general-purpose homogeneous architectures, but there are few added design constraints
- SoC architectures are deployed in many computation environments that require concurrent execution of several tasks with different, sometimes opposite, performance goals
- Integration of different services on the same computing platform requires rethinking of the cores, the memory hierarchy, and the interconnect network
- On these platforms, we need to think not only in terms of tasks and task parallelism, but also services and service guarantees

- Quality of service at the on-chip network level

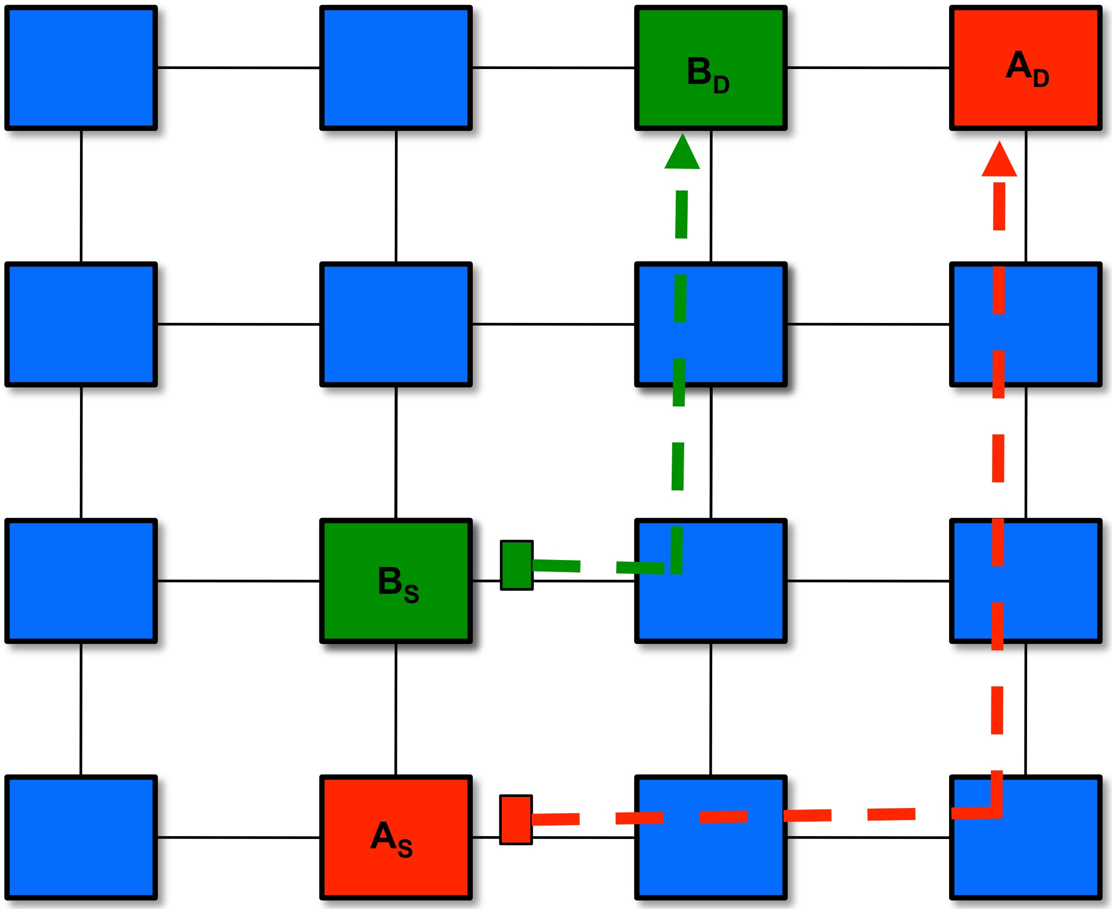
Network Impact on Performance



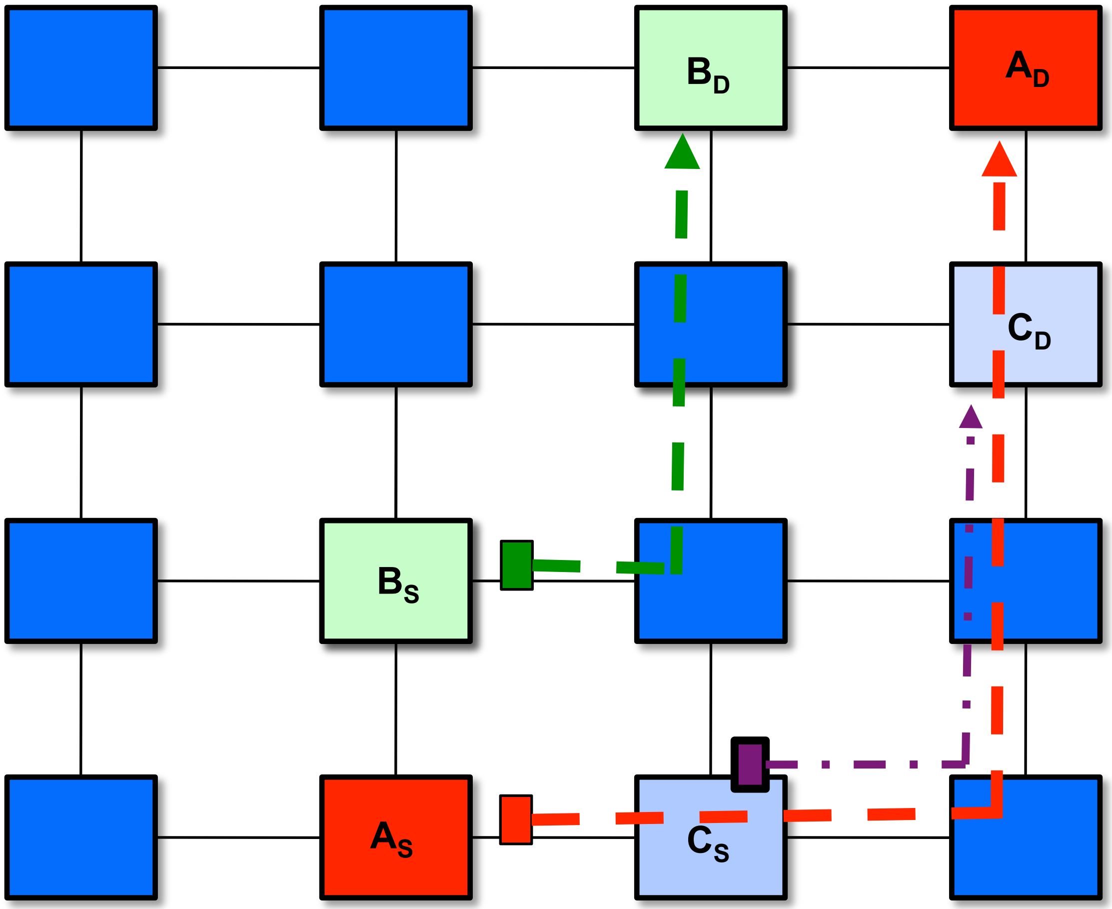
Dynamic Nature of On-Chip Routing



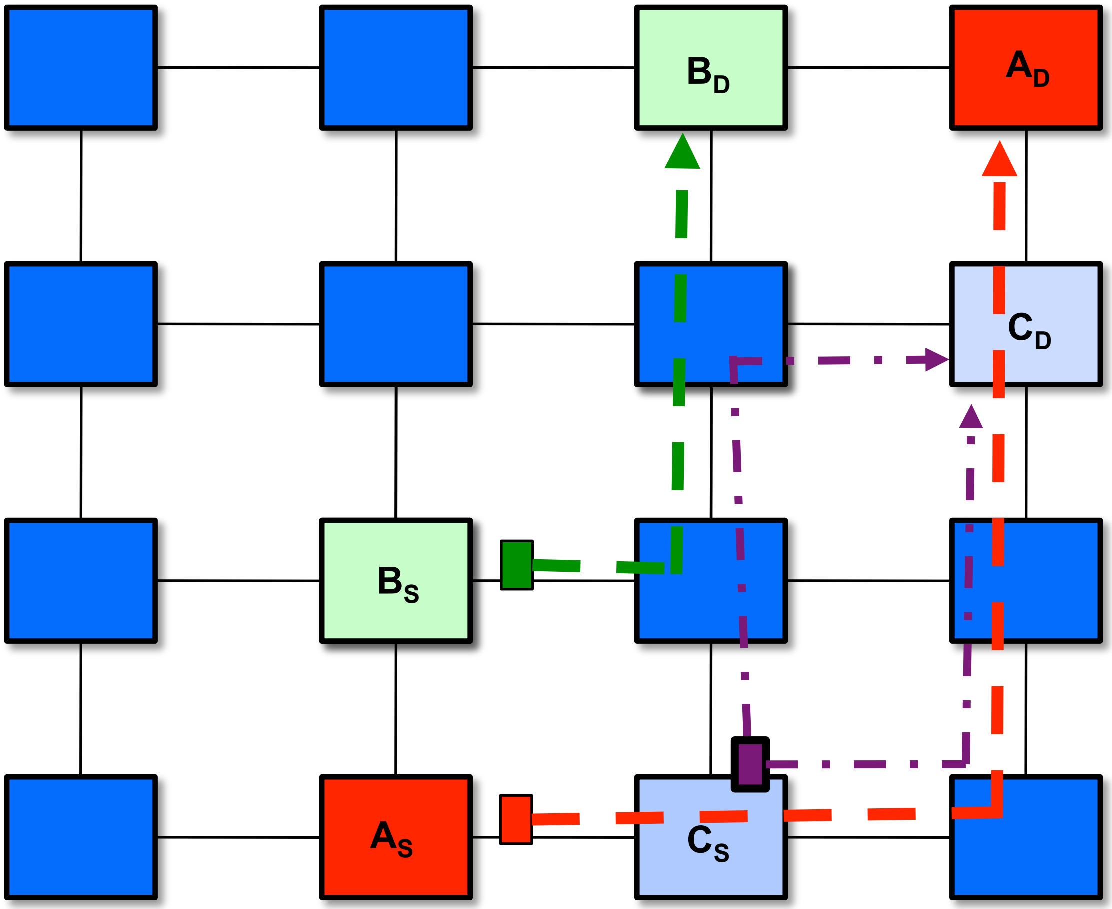
Dynamic Nature of On-Chip Routing



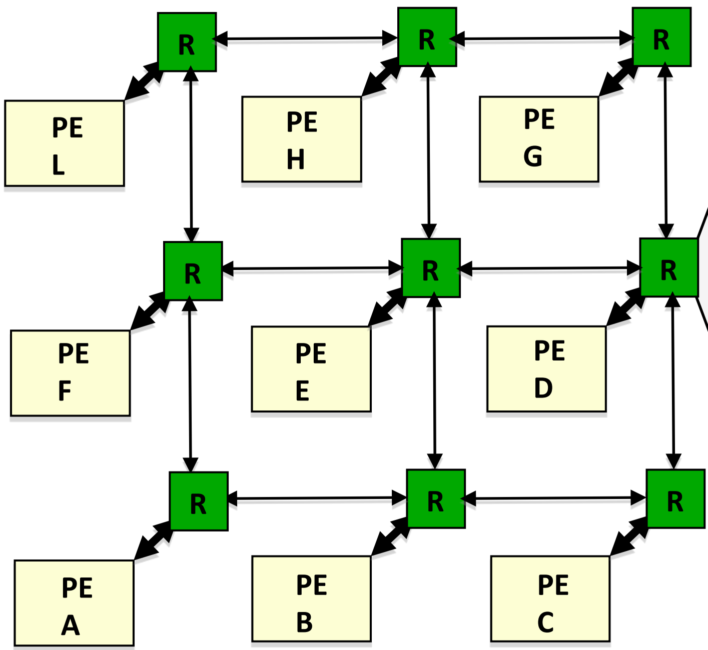
Dynamic Nature of On-Chip Routing



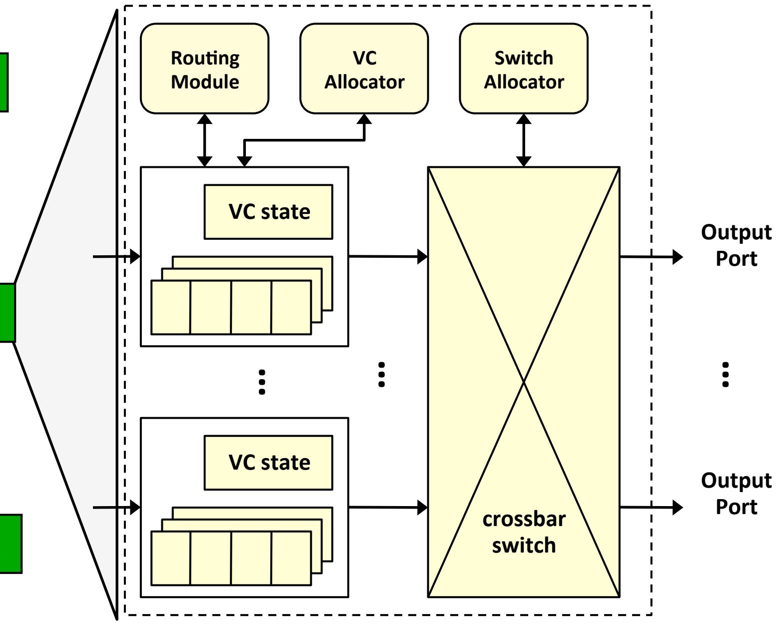
Dynamic Nature of On-Chip Routing



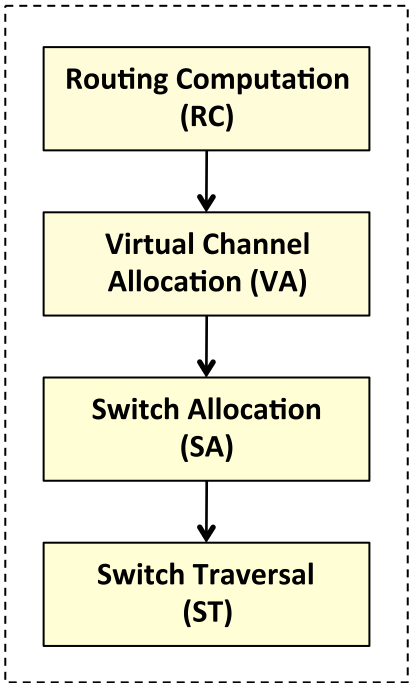
Conventional Router



3x3 Mesh network



Router architecture



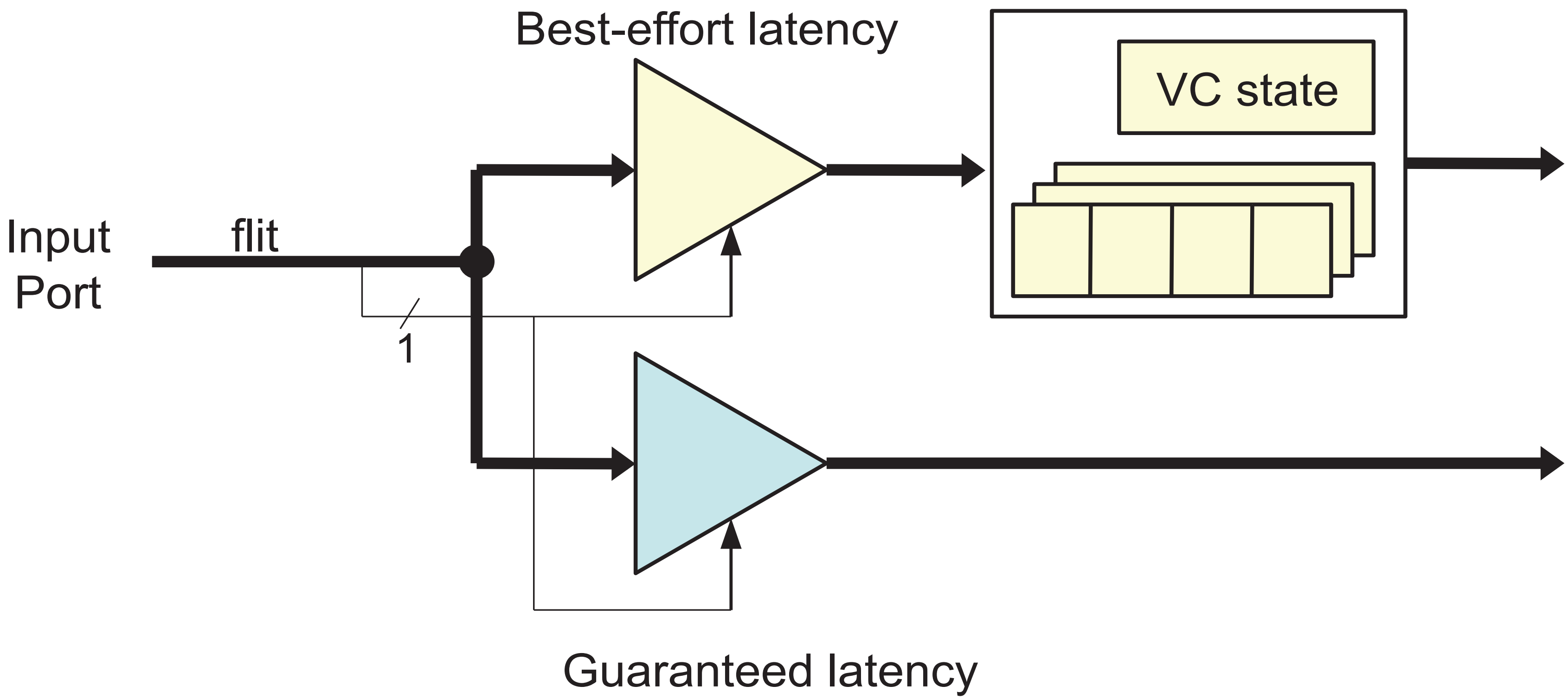
Router: routing phases

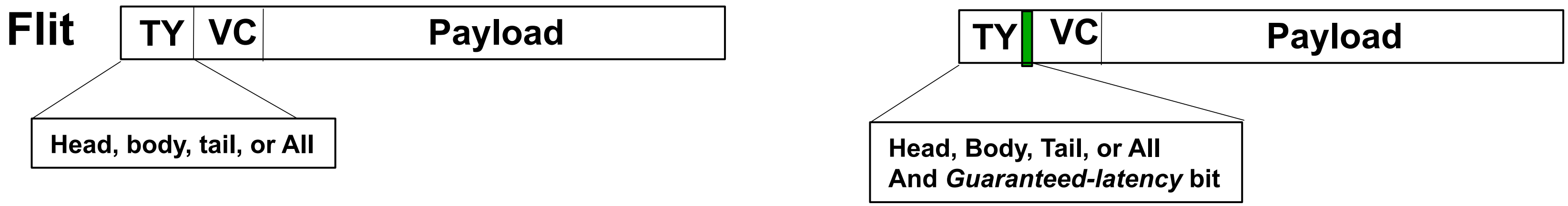
Network Level Quality of Service

The Hard Real-time Support (HRES) router is able to:

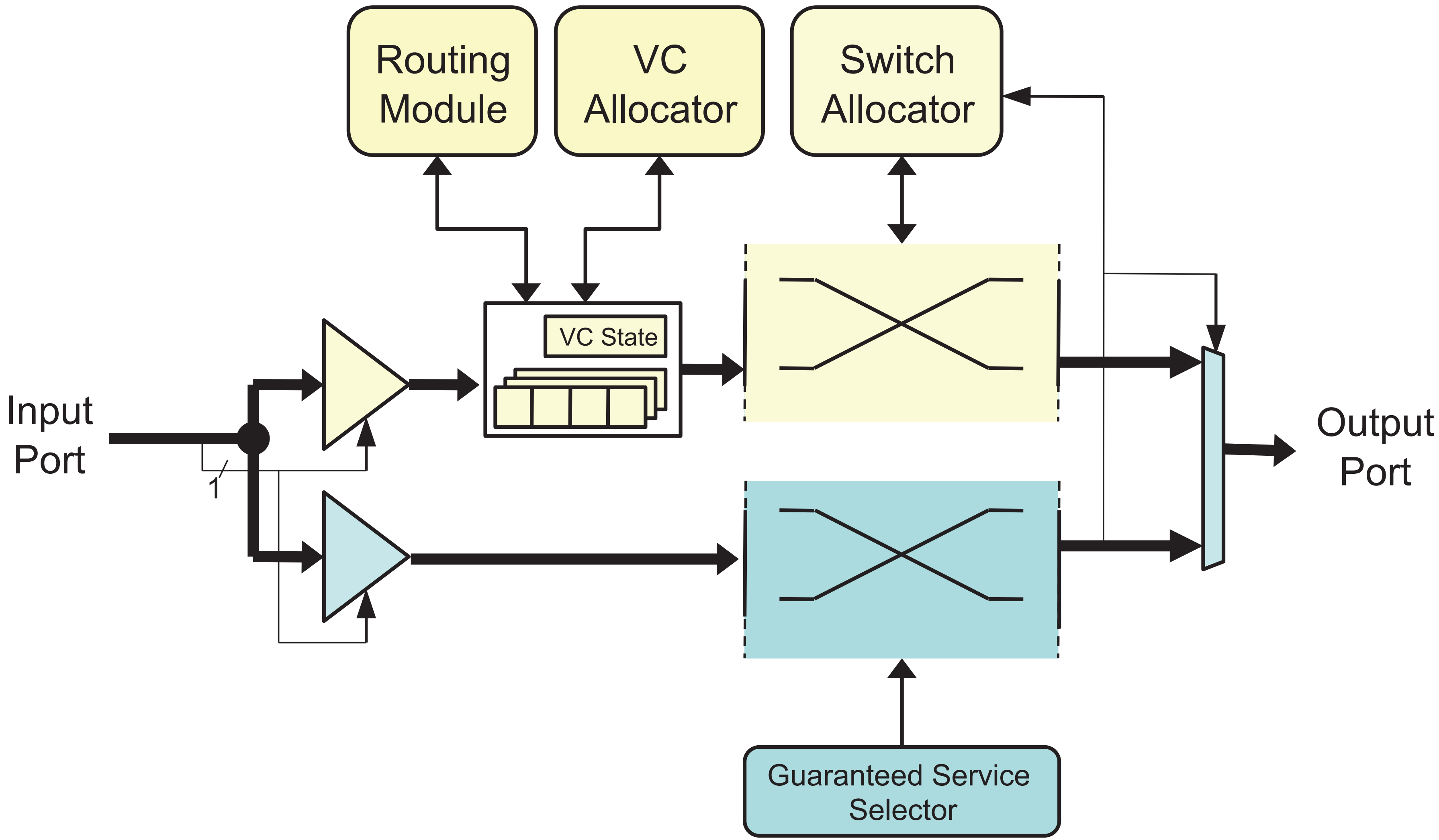
- Decouple hard real-time and best effort traffic using a two-datapath routing scheme
- Maximize link throughput and guarantee hard real-time timing constraints
- Provide fairness of link utilization among the two classes of traffic
- Be acknowledgment-free, retransmission-free, and lossless
- Be deadlock and livelock free with no modification to the buffered datapath of the router.
- It has low hardware overhead for supporting hard real-time communications.

Hard Real-time Support (HRES) router

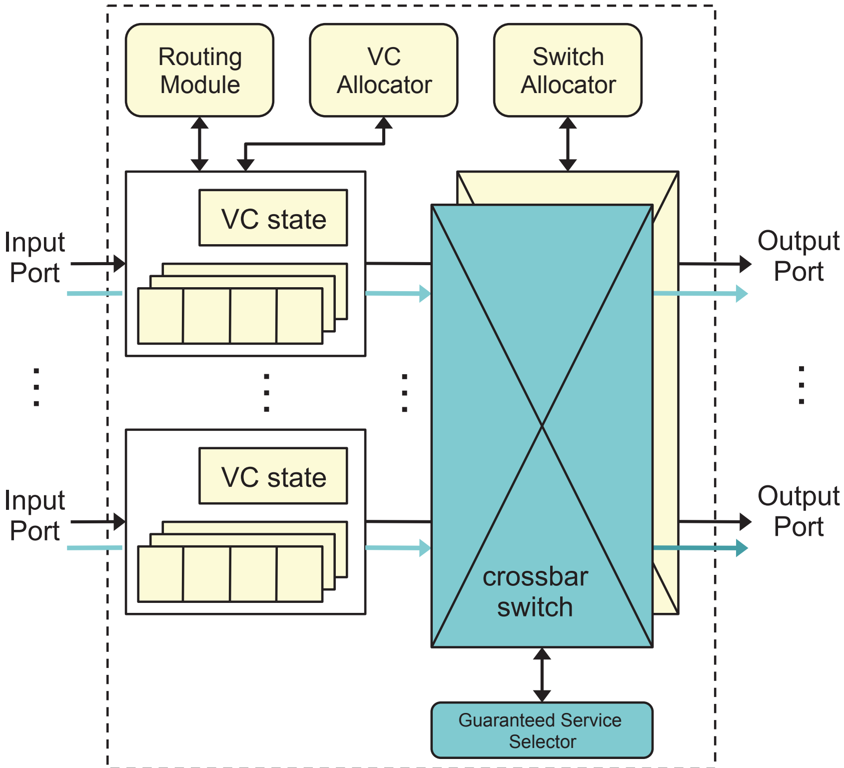




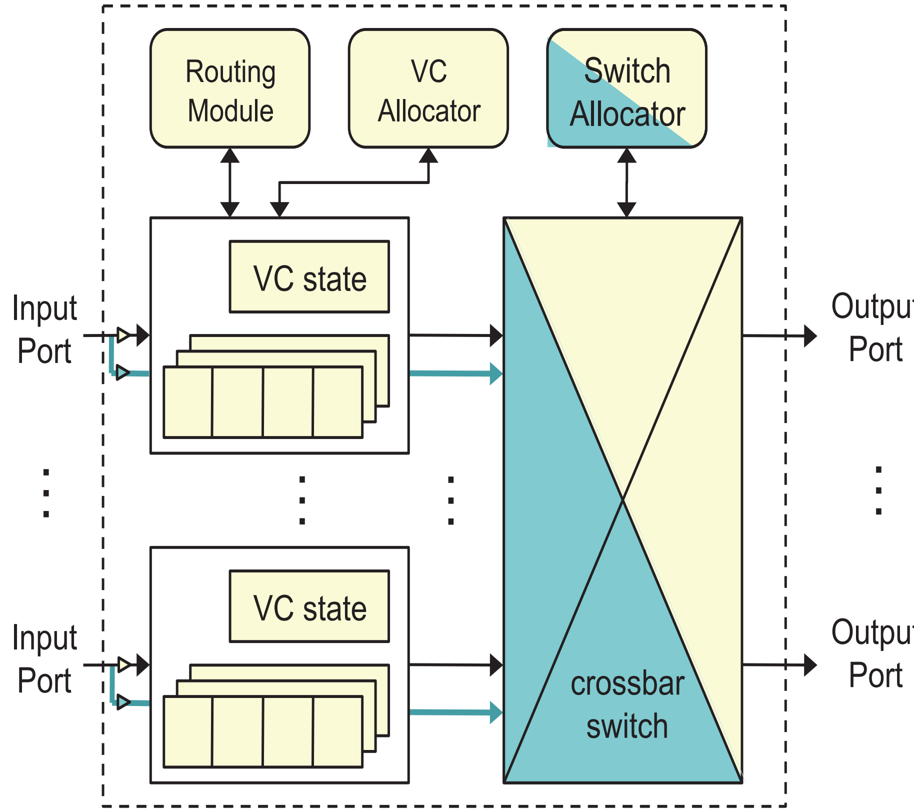
Hard Real-time Support (HRES) router



Other Routers



Two-network router



Single shared crossbar router

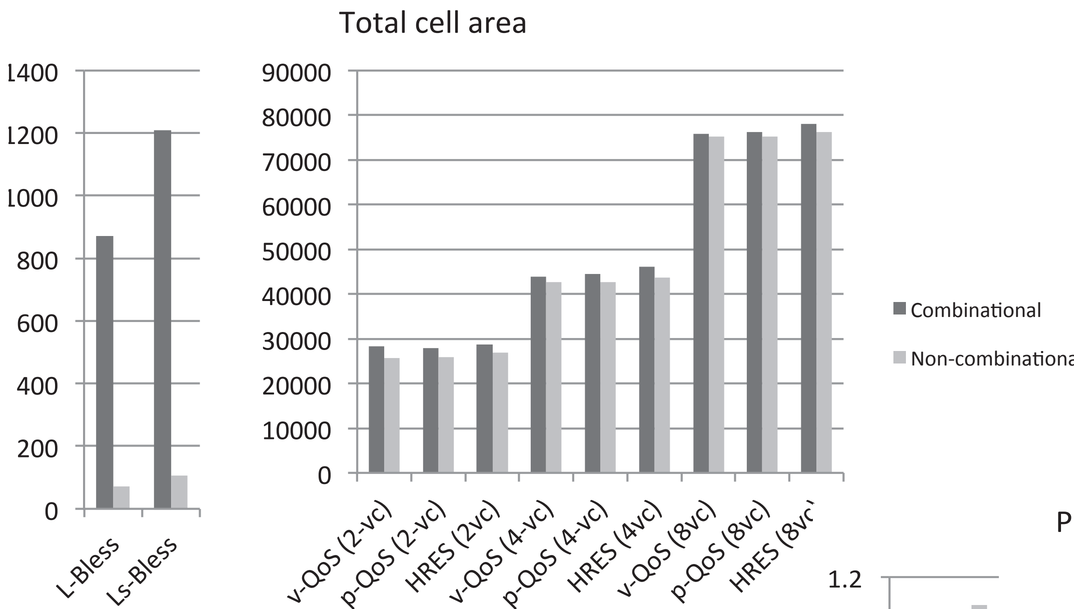
Area Comparison of Architectures

	Number ports	Cell area
HRES router	392	47190.44
Two-Network router	721	51766.84
Single crossbar router	392	52812.94

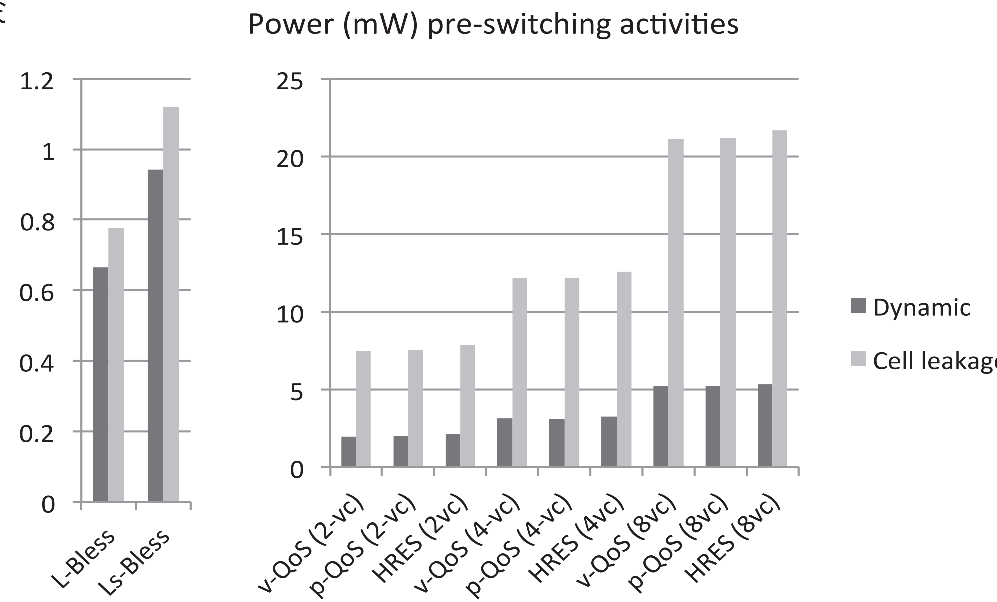
- **Two-network router has duplication of wires and logic**
  - **Lead to more cell area**
  - **Changes in the network interface**

- **Single large crossbar: switch arbitration logic is modified to give priority to the real-time traffic**
  - **Increase in switch arbitration datapath and router critical path**
  - **Real-time and normal traffic requests must be serialized**

Area and Power



- **IBM 45-nm SOI CMOS technology cell library**
- **Comparable static power at the different design points**



- **Comparable logic area at the different design points**

QoS-Aware On-Chip Routing

$$\sum_{i} \sum_{(u,v)\in E} f_{i[r]}(u,v) \cdot w_{i[r]}(u,v) \quad (1)$$

subject to:

$$\forall i,\ \forall (u,v) \in E \quad f_{i[r]}(u,v) \in S_{[r]}(u,v) \quad (2)$$

$$\left|S_{[r]}(u,v)\right| \leq 1 \quad (3)$$

$$0 \leq \sum_{i} f_{i[r]}(u,v) \leq c_{[i]}(u,v) \leq c(u,v) \quad (4)$$

$$\forall (u,v) \in E \quad (\sum_{i} f_i(u,v) + \sum_{i} f_{i[r]}(u,v)) \leq \ c(u,v) \quad (5)$$

$$\forall i \quad f_{i[r]} \in K_{i[r]} \quad (6)$$

$$\left|K_{i[r]}\right| \leq k_i \quad (7)$$

$$\forall i,\ \forall j \in \{1, ..., k_i\} \frac{\sum_{(u,v)\in E} f_{(i,j)[r]}(u,v)}{f_{(i,j)[r]}} \leq deadline_i \quad (8)$$

QoS-Aware On-Chip Routing

$$\sum_{i} \sum_{(u,v)\in E} f_{i[r]}(u,v) \cdot w_{i[r]}(u,v) \tag{1}$$

subject to:

$$\forall i,\ \forall (u,v) \in E \quad f_{i[r]}(u,v) \in S_{[r]}(u,v) \tag{2}$$

$$\left|S_{[r]}(u,v)\right| \leq 1 \tag{3}$$

Non-starvation

$$\boxed{0 \leq \sum_{i} f_{i[r]}(u,v) \leq c_{[i]}(u,v) \leq c(u,v)} \tag{4}$$

$$\forall (u,v) \in E \quad \left(\sum_{i} f_{i}(u,v) + \sum_{i} f_{i[r]}(u,v)\right) \leq \ c(u,v) \tag{5}$$

$$\forall i \quad f_{i[r]} \in K_{i[r]} \tag{6}$$

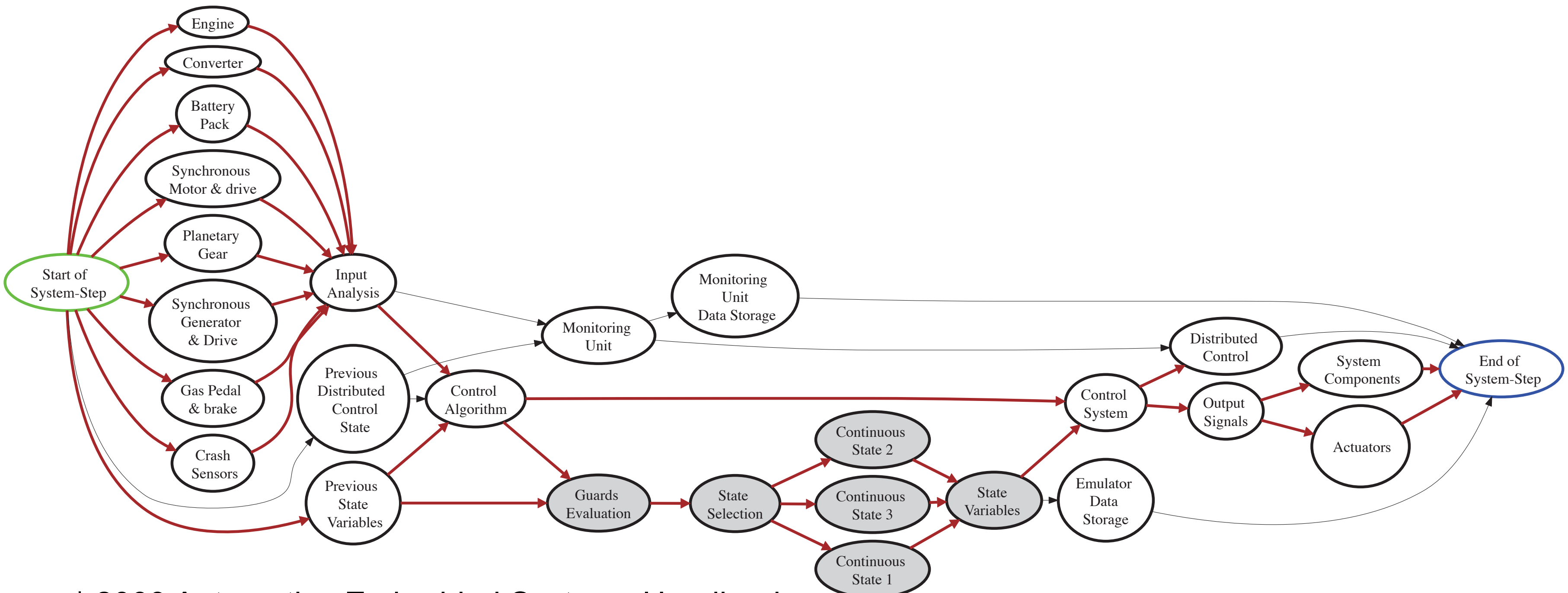
$$\left|K_{i[r]}\right| \leq k_i \tag{7}$$

Deadlines

$$\boxed{\forall i,\ \forall j \in \{1, \ldots, k_i\} \frac{\sum_{(u,v)\in E} f_{(i,j)[r]}(u,v)}{f_{(i,j)[r]}} \leq deadline_i} \tag{8}$$

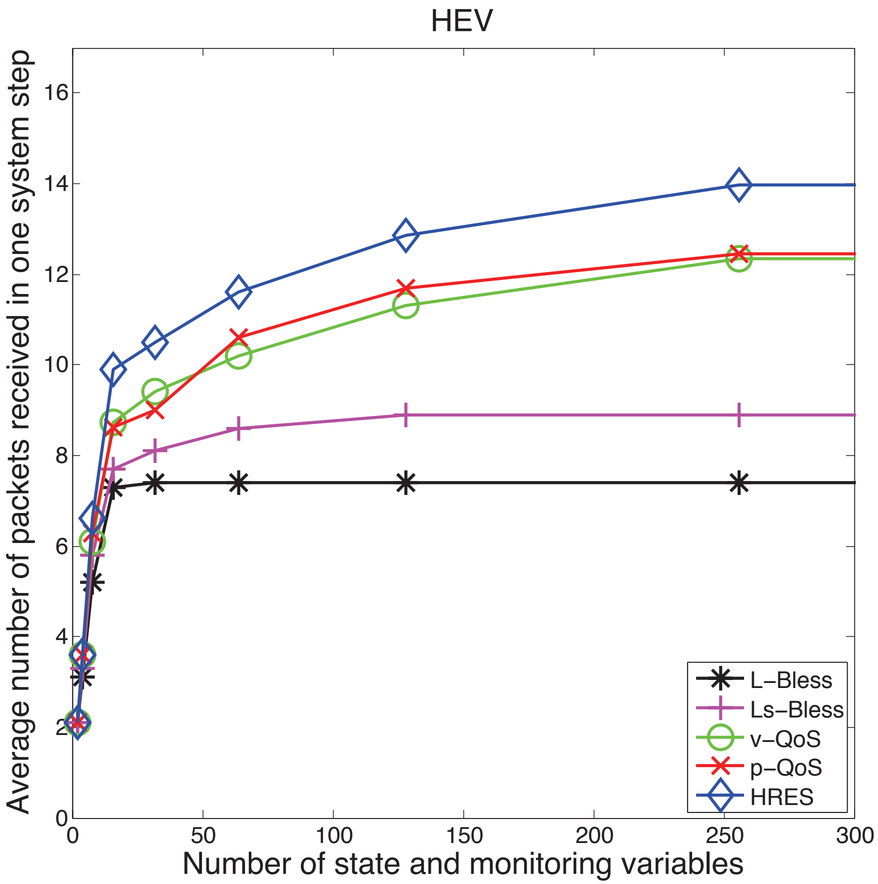
Hybrid electric vehicle (HEV) Application

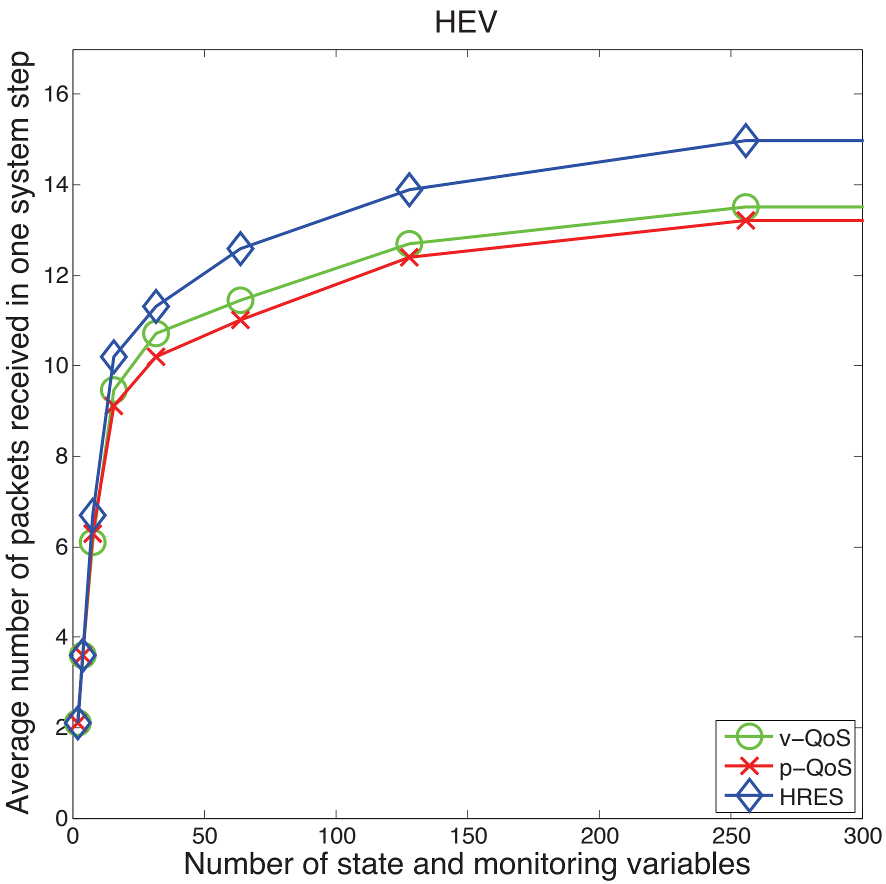
- **In a modern automobile there are as many as as 70 electronic control units (ECUs) embedded in a vehicle***
- **In the HEV, a motor drive, with a controlled inverter system, is needed to deliver powerful and efficient drive to the electric motor**
- **Mixed-criticality application: hard real-time and best effort**



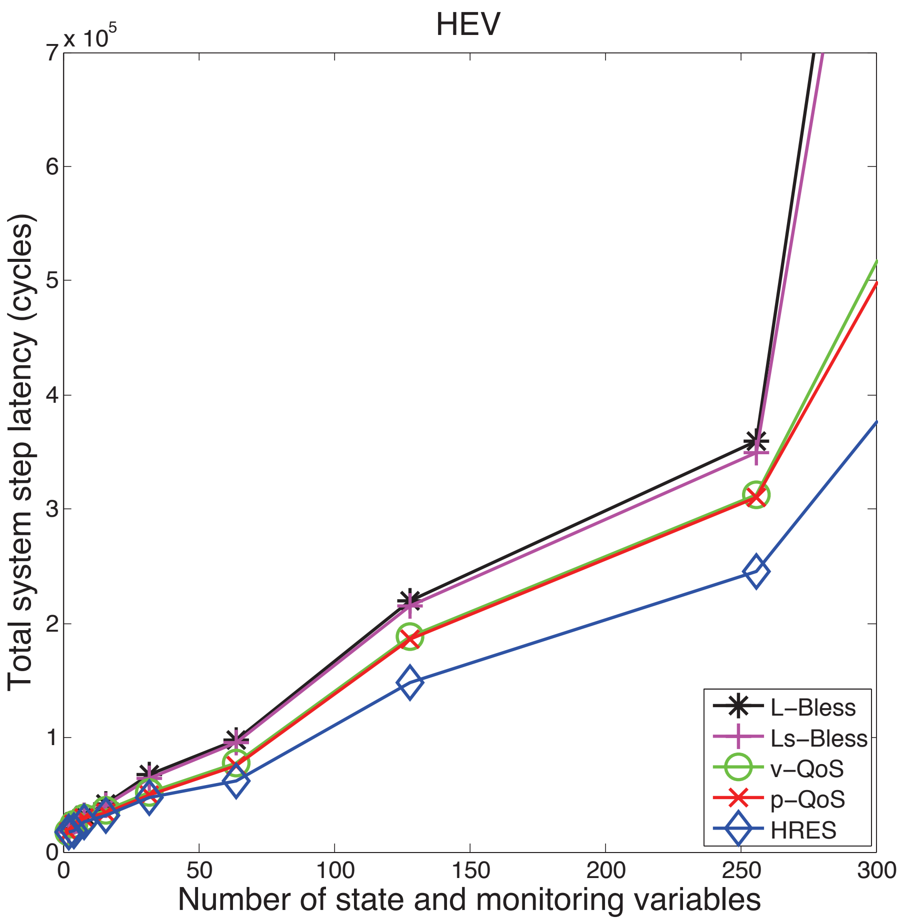
* 2009 Automotive Embedded Systems Handbook

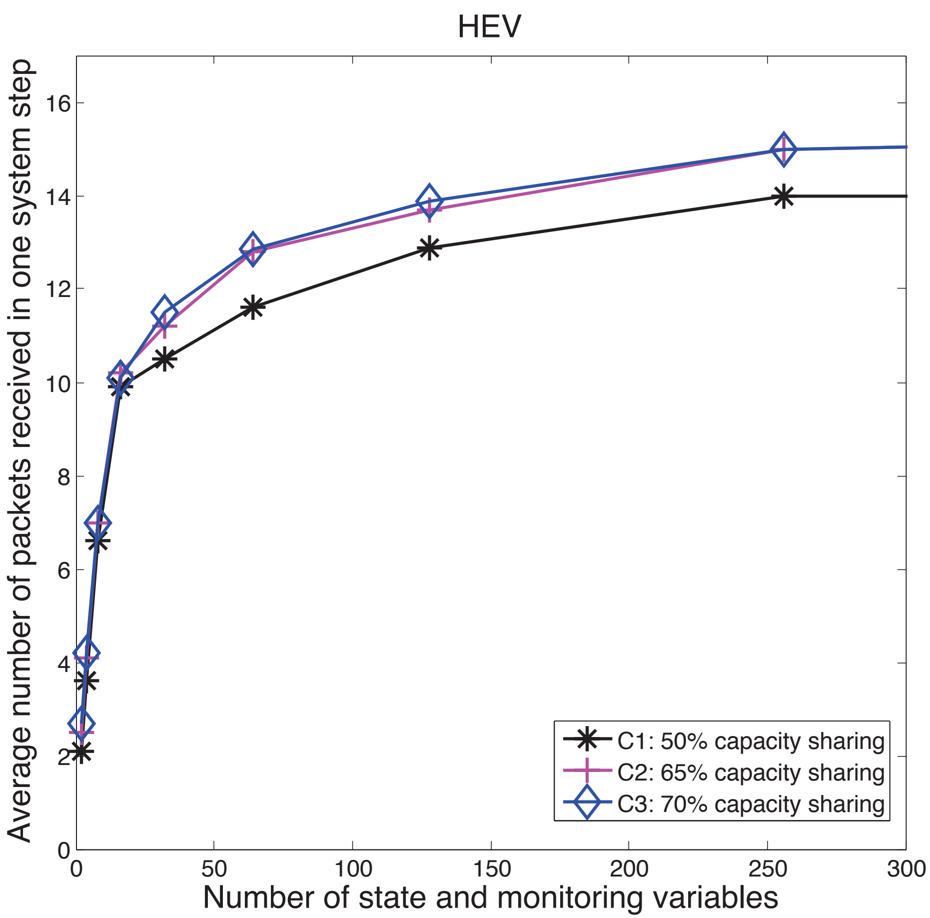
Evaluation Results





Evaluation Results





Q & A Section

Thank You !

More Information at
http://caes.cs.uoregon.edu/