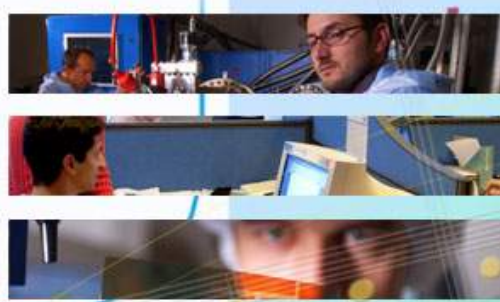
micro and nanoelectronics
microsystems
ambient intelligence
image chain
biology and health

2008

Automatic Power Regulation based on an Asynchronous Activity Detection and its Application to ANOC Node Leakage Reduction

Yvain Thonnart
Edith Beigné
Alexandre Valentian
Pascal Vivet

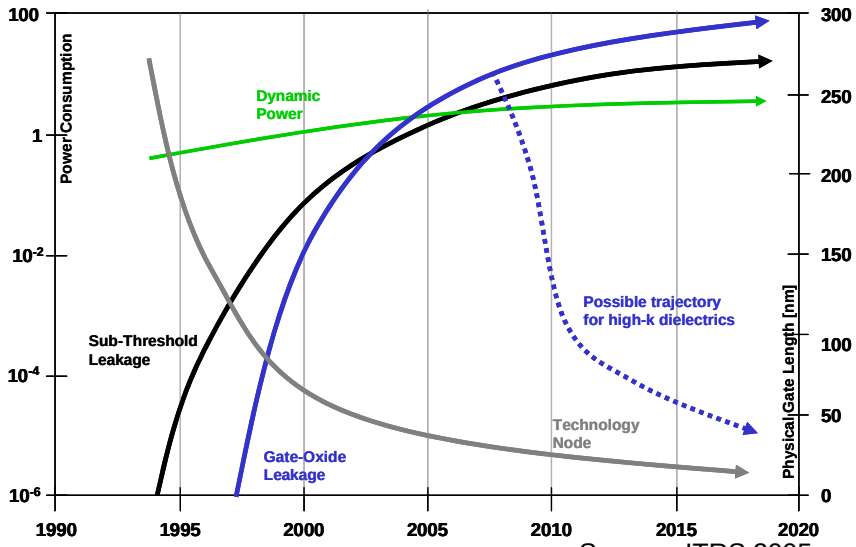


leti MINATEC INSTITUT CARNOT CEA LETI

cea

Leakage issue in complex SoCs

- ITRS trends:
  - "Power consumption is now the major technical problem facing the semiconductor industry"
  - "Leakage power (subthreshold, gate) increases exponentially as process moves to finer technologies"



Source : ITRS 2005

- What about asynchronous circuits ?
  - ☺ dynamic power : reduced compared to clocked designs
  - ☹ static power : suffers from a larger area (especially for QDI circuits)

Table DESN4a Logical/Circuit/Physical Design Technology Requirements—Near-term Years

Year of Production	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*
Asynchronous global signaling: % of a design driven by handshake clocking	7%	11%	15%	17%	19%	20%	22%	23%	25%
Full-chip leakage (normalized to full-chip leakage power dissipation in 2007)	1	1.5	2	2.5	2.75	3	3.5	4	6

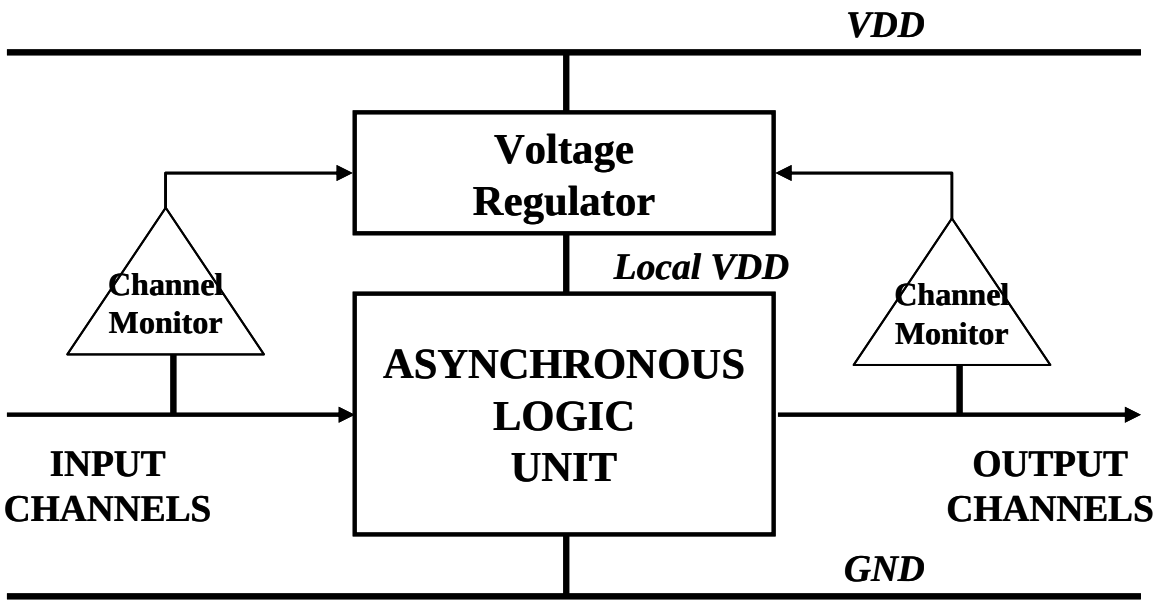
Source : ITRS 2007 - Design

Outline

- A pragmatic approach : from small to big designs
  - Leakage issue in asynchronous circuits
  - Asynchronous Activity detection scheme
  - Local handshake monitoring and idle mode
  - Higher level monitoring of I/Os and low-power mode
- Case study on ANOC
  - Activity detection architecture
  - Power regulation design
- Results on ANOC
  - Physical implementation and area overhead
  - A compromise between performance and energy
- Conclusion and Perspectives
  - Generalization of the method

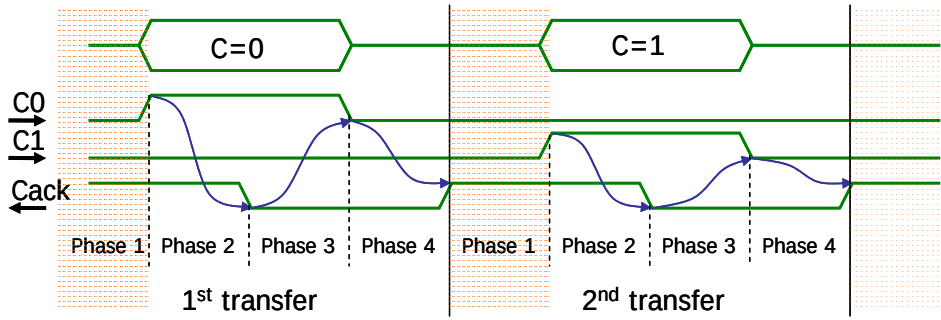
Asynchronous activity detection scheme

- Objectives
  - Detect inactivity on asynchronous channels
  - Place design in idle/standby mode
- Power strategies
  - Power off & global reset
  - State retention (no current injection)
  - Slow switching

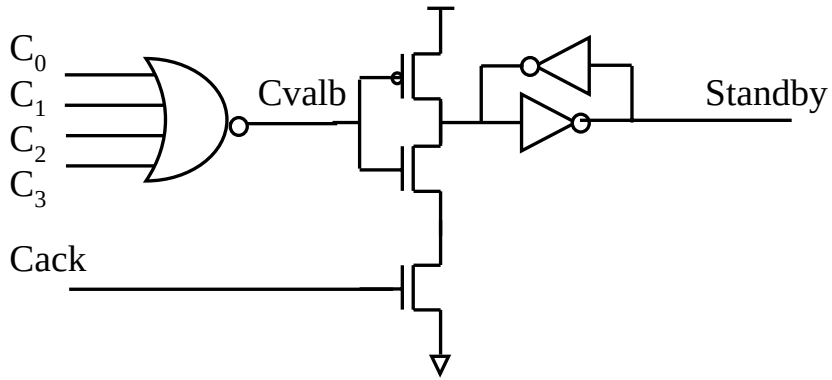


Activity detection at handshake level

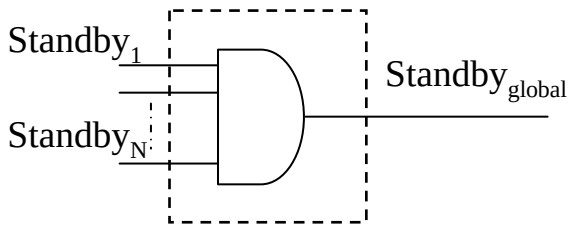
- Principles
  - Monitoring of every asynchronous channel



- Detection
  - Generalized C-Element
  - Set during phase 1

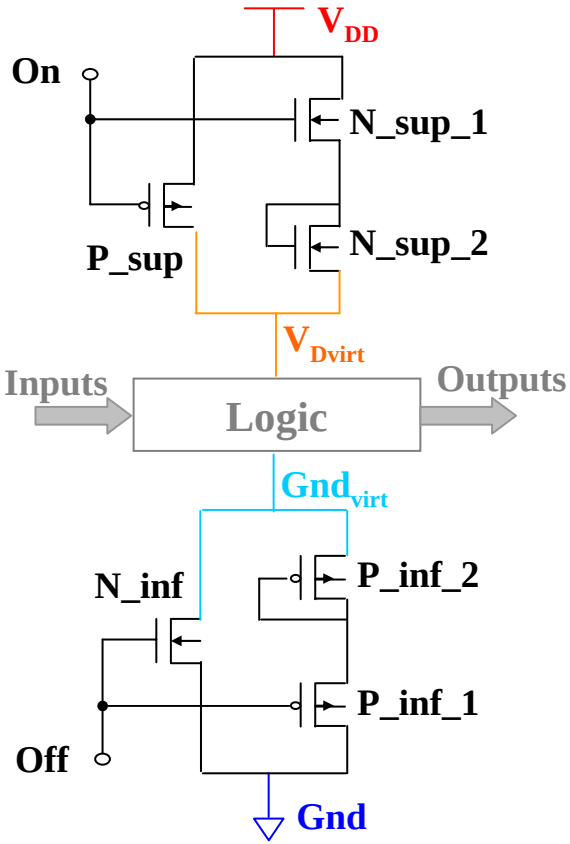


- Collection
  - AND of local signals



Self-controllable voltage level circuit

- Taken from state of the art
  - T. Enomoto, Y. Oka, H. Shikano, “A self-controllable voltage level (SVL) circuit and it’s low-power high-speed CMOS circuit applications”. IEEE Journal of Solid-State Circuits, vol 38, n° 7, jul. 2003, pp. 1220-1226
- Objectives
  - MTCMOS with state retention
  - Avoid the need of a reset at power-on
- Principle
  - Use weakly-on transistors in standby mode to control virtual Vdd and Gnd
- No current is delivered
  - Circuit is stalled
  - Leakage reduction is consequent (2 decades)



Handshake detection example

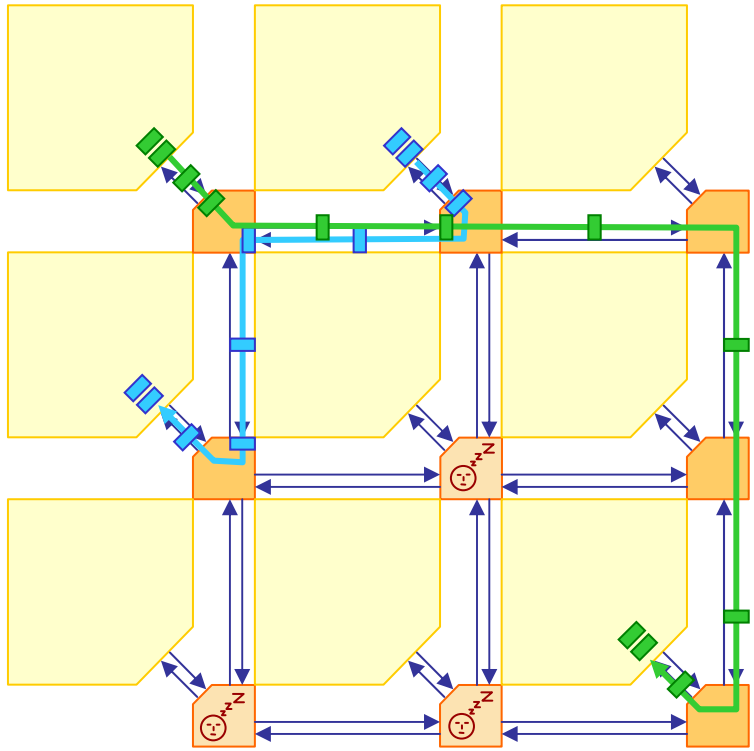
- Case study : pipelined 32 bit ripple carry adder
  - 129 asynchronous channels to monitor in theory
  - Only 34 thanks to the pipeline structure in practice: monitoring of the carry-in and the outputs
- Simulation results
  - Leakage reduction: -98%
  - Speed degradation: -8%
  - Area penalty: +13%
- Limitations for complex designs
  - Monitoring every channel is quickly prohibitive
  - The circuit is inactive during standby
  - Specific power domains are needed for monitoring logic
    - Power stripes, level shifters in deeply embedded logic

Seeing bigger…

- New objectives for complex designs
  - Functional low-power mode
    - Instead of state retention
    - Ability to handle single tokens or short bursts without needing to power-on / power-off
    - No need to control every channel of a bus
  - Monitoring only at Inputs and Outputs
    - Too many channels in a consequent design
  - Non-elementary activity control
    - Processing does not necessarily depend on a single token

Application to ANOC

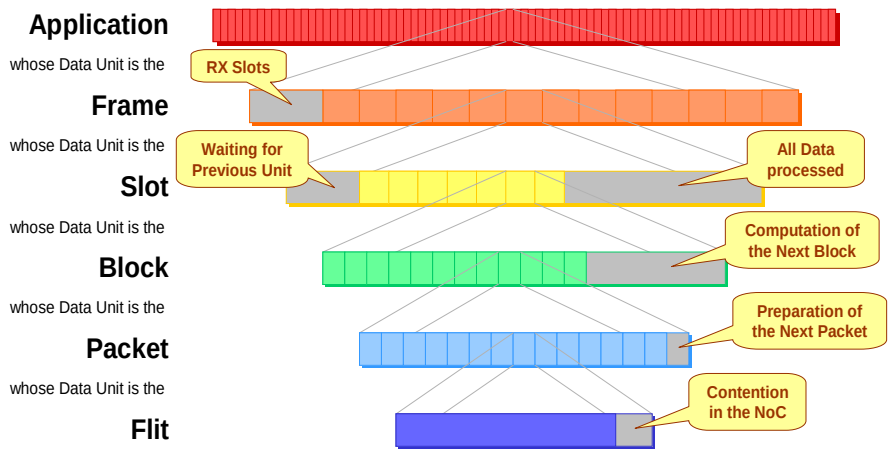
- ANOC architecture
  - QDI asynchronous routers and links
  - Wormhole packed-based communications
- Power analysis (130nm)
  - Leakage proportion during activity : ~25%
  - Some routers are not used during long idle phases
- Proposal
  - Local power-down of each router
  - No wake-up for infrequent signaling messages
  - Power control information embedded in the data packets

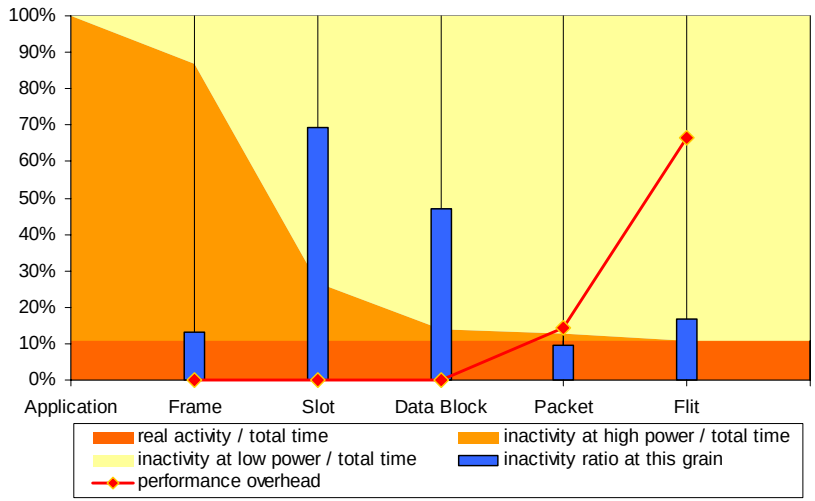


Levels of granularity in NoC traffic

Telecom applications

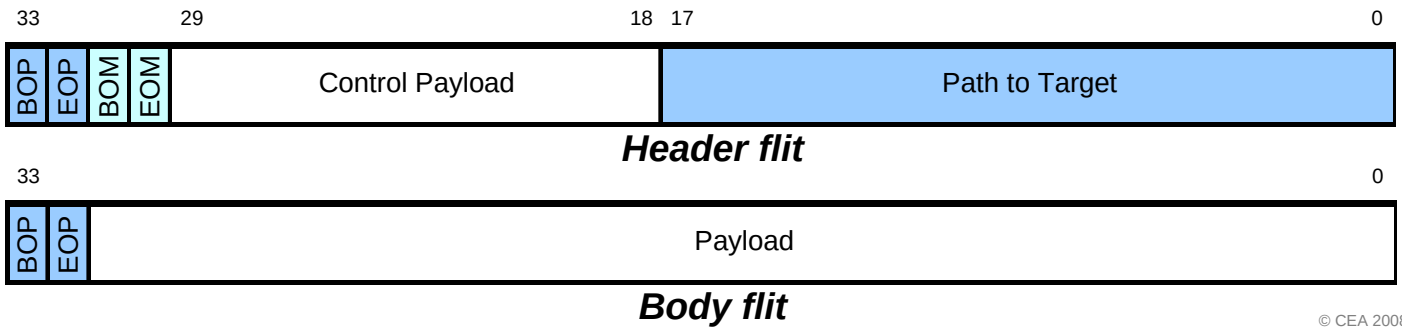
- Temporal repartition :
  - traffic decomposes in several levels
  - Inactivity phases occur at each level
  - Total inactivity may represent up to 90% of the time
- Spatial repartition :
  - Some routers may be unused for long phases (ex : half duplex TX/RX, up to 0.5 ms inactivity)
- Most of the leakage when the handset is not in use





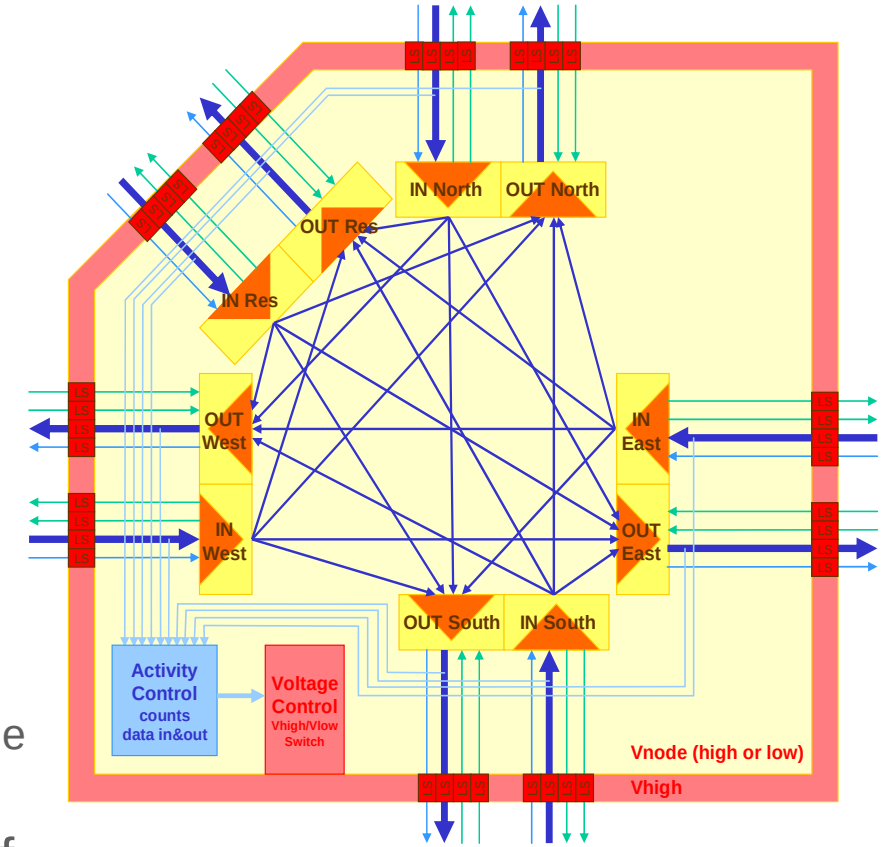
Activity protocol encoding

- Tag packets with activity markers
  - Allows to select whichever granularity level from packet to application
  - Begin of Message
    - request power-on of all routers on the path of the packet
  - End of Message
    - clear request : routers will power-down when no more request is active
- Packets tagged :
  - Hopefully, BOM & EOM are also used in Network Interfaces :
    → begin & end of packet bursts
  - DATA (streaming)
  - CREDIT (streaming flow control)
  - MOVE (memory transfers)
  - Other signalling packets (IT, …) do not affect activity



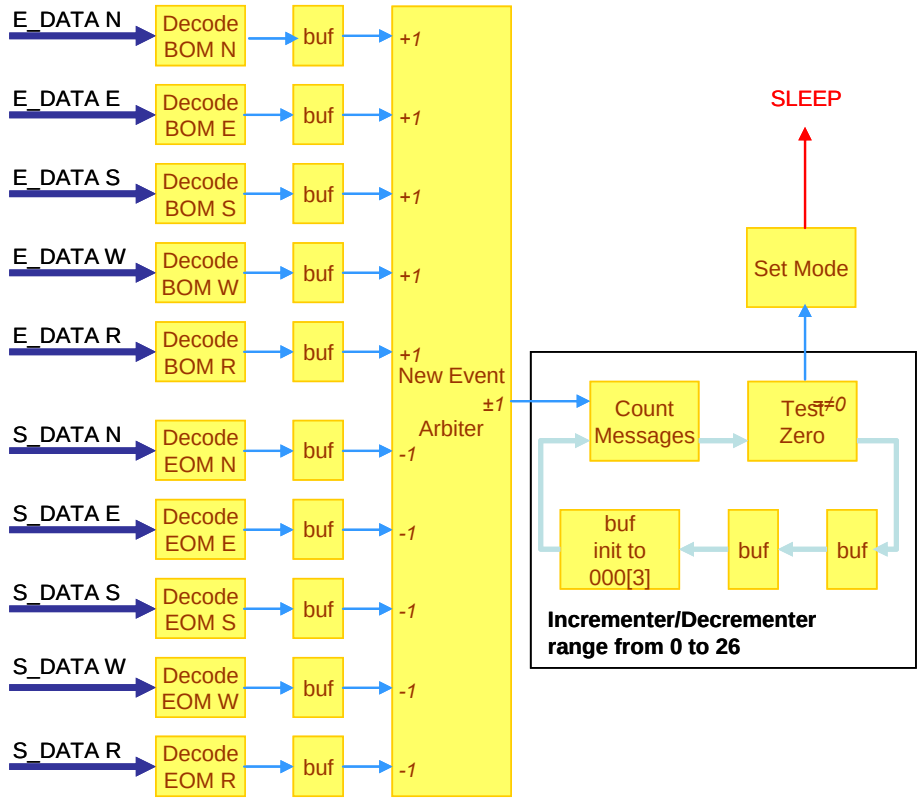
Router architecture

- Activity detection
  - QDI I/O monitoring
    - Fork on forward paths
    - C-element on ack. paths
  - Concurrent flows counter
    - BOM on an input: +1
    - EOM on an output: -1
- Power planning
  - Power switch
    - To regulate core voltage according to activity
  - Level shifters on all I/Os
    - To maintain nominal voltage on network links
- Core voltage never turned off
  - No reset generation logic



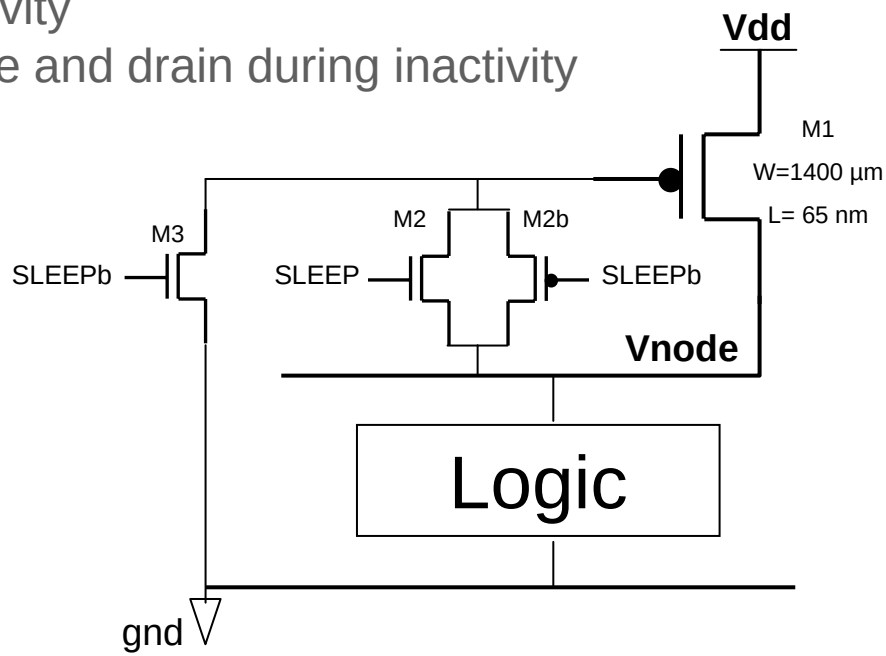
Activity detection micro-architecture

- Principles
  - Buffers on I/Os
    - To keep the channels free
  - Arbiter on BOM/EOMs
    - Handle each request/release successively
  - Incrementer/Decrementer
    - « inverted semaphor »
    - Check if shared resource is not used at all
- Overflow handling
  - Counter sizing for 5 flows/input
    - From applicative observations (2-3 max)
  - No coherence loss
    - Temporary extra low-power only



Power regulation design

- Principles
  - Main M1 PMOS power switch
  - M3 closes the switch during activity
  - Pass-Gate M2 connects M1 gate and drain during inactivity



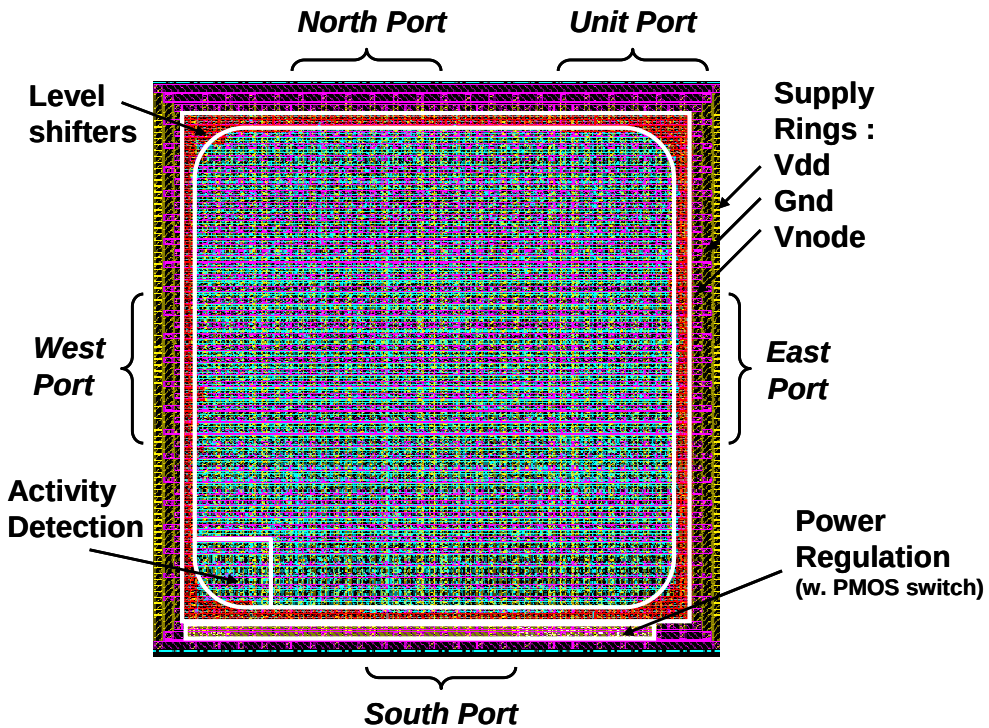
- Sizing of M1
  - According to max dynamic current from simulation
  - To guarantee less than 5% delay penalty during activity
  - Vdd * 1/2 < Vnode < Vdd * 2/3

- Current may be delivered in low-power mode
  - Logic is functional
  - Speed is reduced

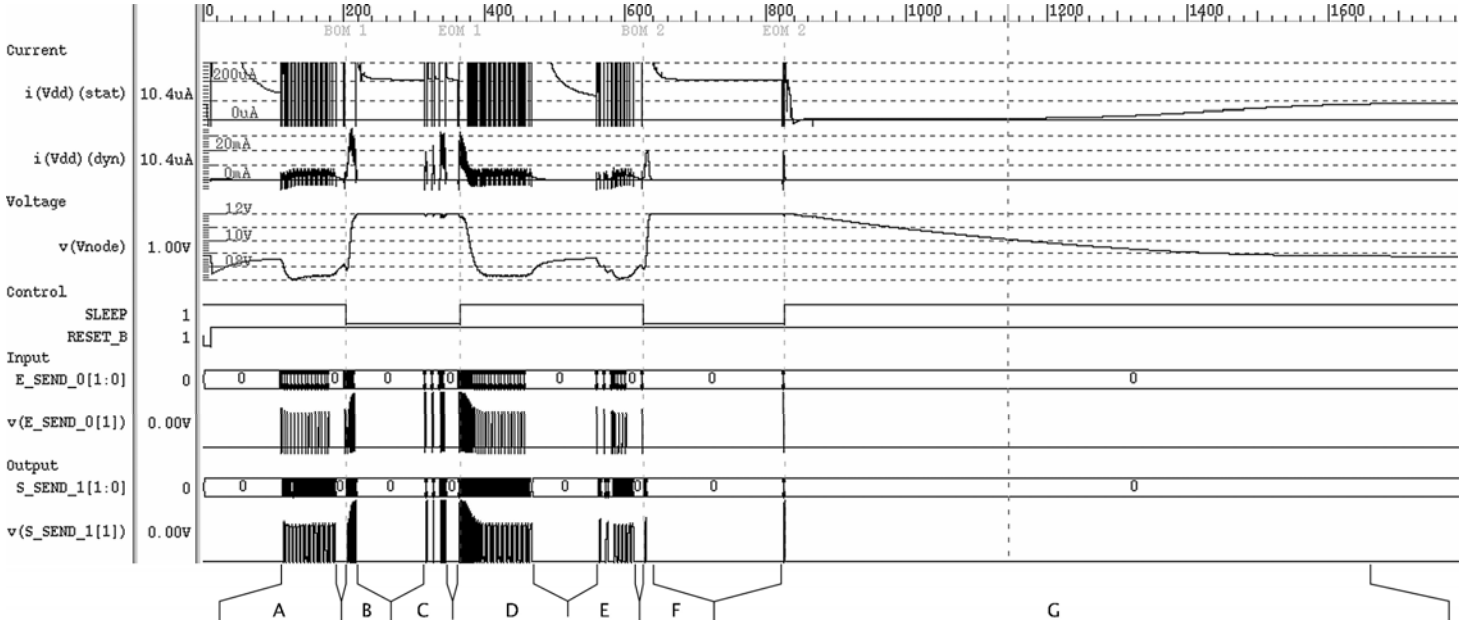
Physical implementation

- Implementation of the router
  - STMicroelectronics CMOS 65 nm technology
  - Integrated in the ALPIN chip (6 power-aware routers)
  - Area overhead
    - Mostly due to the level shifters
    - The switch is under the power rings

Block	Area
Core	85%
Level-Shifters	13%
Activity Detection	2%
Power Switch	0%
Total	200,000 µm²



Electrical simulation



- Voltage never drops under 0.7 V
- Power-on transitions are fast
- Power-down transitions depend on the load
  - Either fast due to dynamic switching
  - Or slow under no load
    - Precisely because the leakage currents were reduced

Overall Results

Intrinsic performance

- Almost no degradation at power-on
- 60% leakage reduction
- Speed/Energy tradeoff at power-down
  - ¼ Speed
  - ½ Energy

	Original version	Power-On	Power-down
Supply voltage	1.2V	1.2V	0.6-0.8V
Flit Throughput	1.8 ns (550Mflit/s)	1.8 ns (550Mflit/s)	7.2 ns (140Mflit/s)
Flit Latency	2.3 ns	2.5 ns	5.8 ns
Leakage	200 µA (240µW)	210 µA (250µW)	80 µA (100µW)
Energy	30 pJ/flit	30 pJ/flit	14 pJ/flit

Applicative performance

- Based on a telecom chain
- Router at FFT output
- From 10% to 60% total power consumption reduction
- 60% on idle NoC routers

Conditions MIMO MC-CDMA traffic	Dynamic Power	Static Power	Total Power	Gain
Always at power-on	1 mW	250 µW	1.25 mW	-
Auto-detection	1 mW	120 µW	1.12 mW	10 %
Always at power-down	0,7 mW	100 µW	0.80 mW	36 %
Idle at power-on	0 mW	250 µW	250 µW	-
Idle at power-down	0 mW	100 µW	100 µW	60 %

Conclusion

- A technique to reduce leakage during inactivity in asynchronous circuits
- Applied to an asynchronous network on chip
- Also allows for dynamic savings when speed is not crucial
- Generalization
  - Token counting activity detection
    - Applicable to any slave resource (shared or not)
      - Identify the Begin of a request / End of a response
  - Power regulator
    - Applicable to any QDI circuit
    - Should be sized according to:
      - Maximum consumption
      - Real-time constraints at high and low power

micro and nanoelectronics
microsystems
ambient intelligence
biology and health
image chain

Innovation for industry

Loyalty
Entrepreneurship
Team work
Innovation

leti

MINATEC

INSTITUT CARNOT CEA LETI

cea