An 800MHz Star-Connected On-Chip Network for Application to Systems on a Chip

Se-Joong Lee, Seong-Jun Song, Kangmin Lee,
Jeong-Ho Woo, Sung-Eun Kim,
Byeong-Gyu Nam, and Hoi-Jun Yoo
Semiconductor System Lab., KAIST, Korea

Presentation Outline

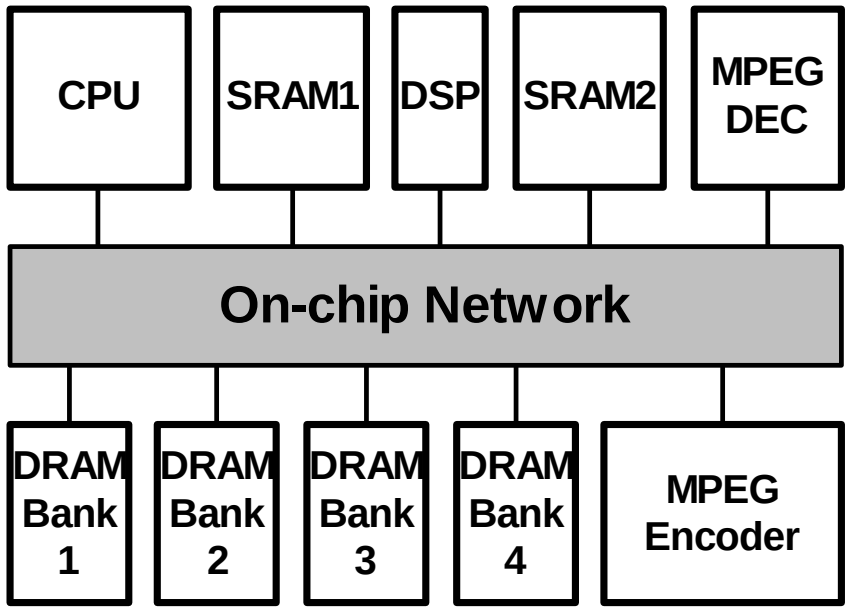
- **Motivation**
- **Techniques for**
  - **Bandwidth guarantee**
  - **Latency minimization**
  - **Area reduction**
- **Special Features**
  - **Plesiochronous communication**
  - **Chip-chip transaction**
- **Measurement Results**
- **Conclusion**

Motivation

- **Issues on SOC**
  - **Number of IPs**
  - **Bandwidth bottleneck**
  - **Global synchronization**

⇩

- **On-chip Network (OCN)**
  - **System integration platform**
  - **Networking without global synchronization**

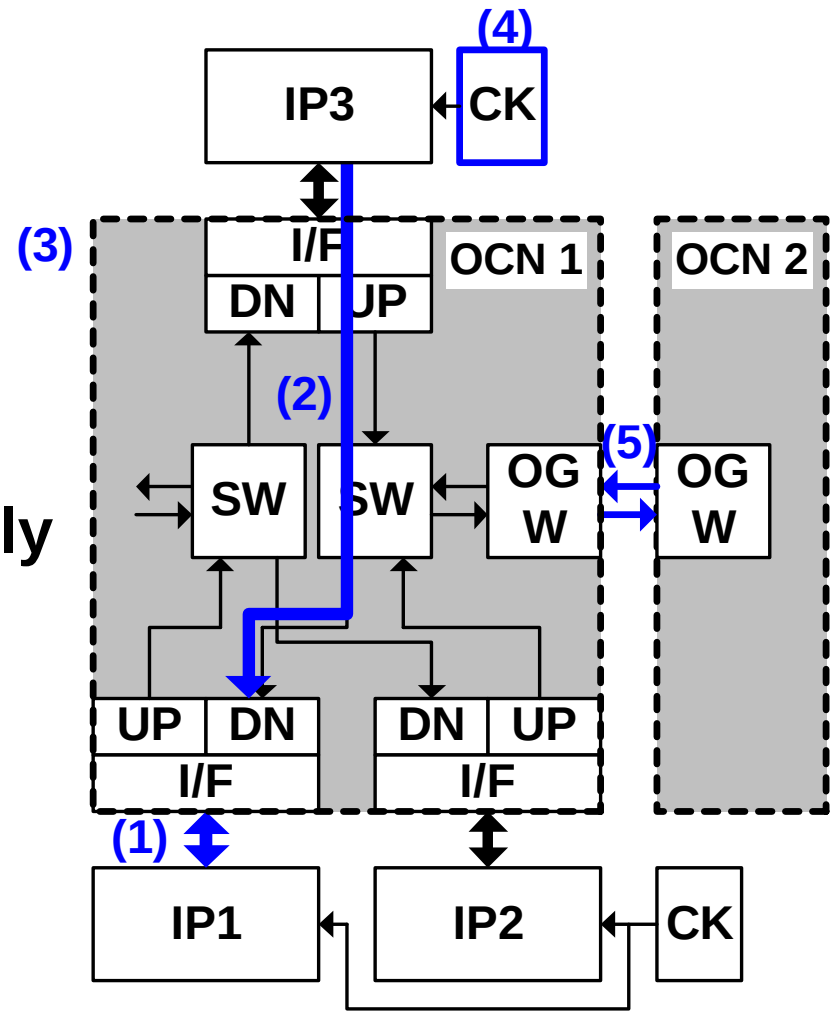


- **This Work**

  → **First Implementation of On-chip Network**

Requirements of SOC

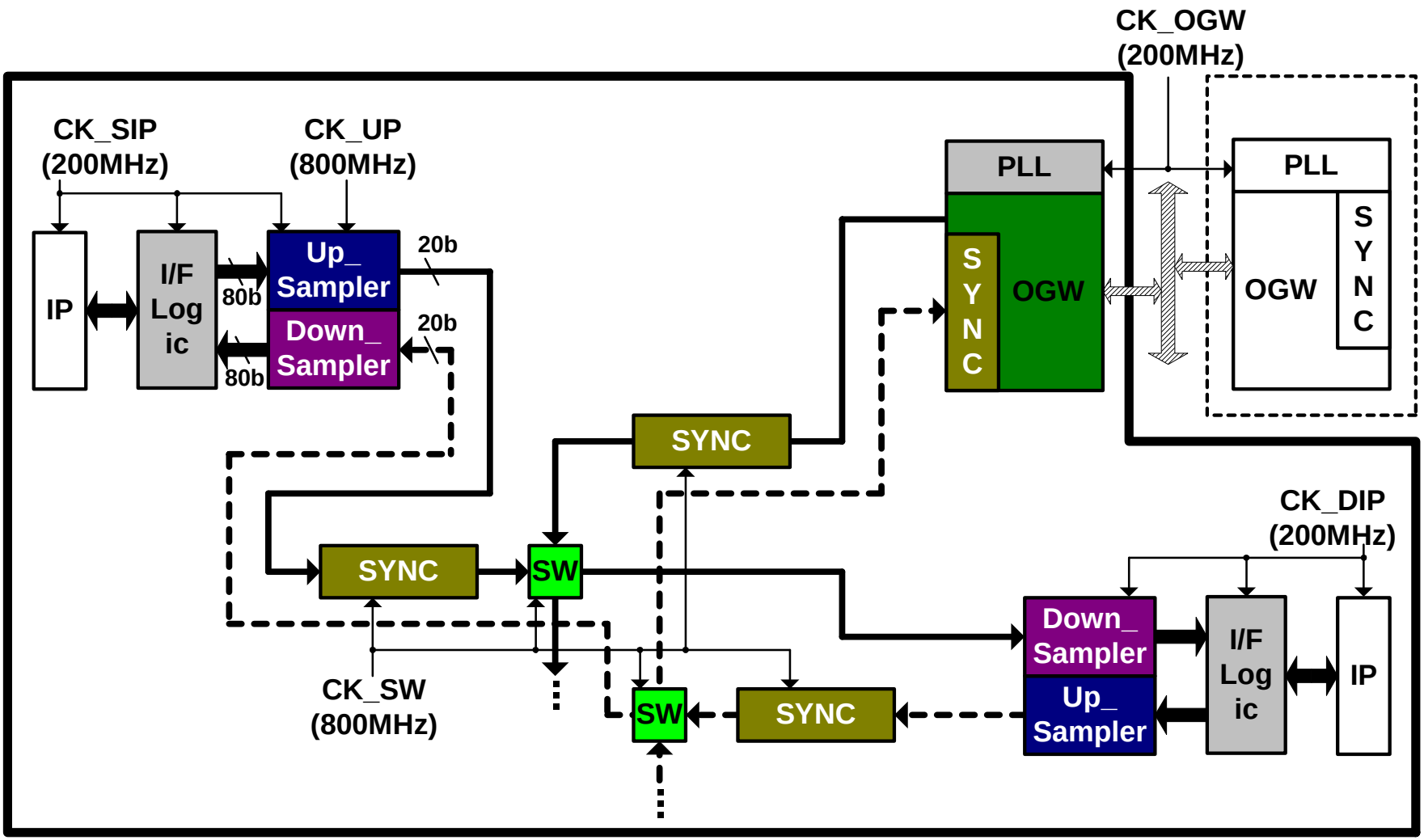
1) **Full bandwidth**
2) **Deterministic latency**
3) **Low area overhead**
4) **Globally asynchronous locally synchronous**
5) **Chip-to-chip communication**



Features of OCN

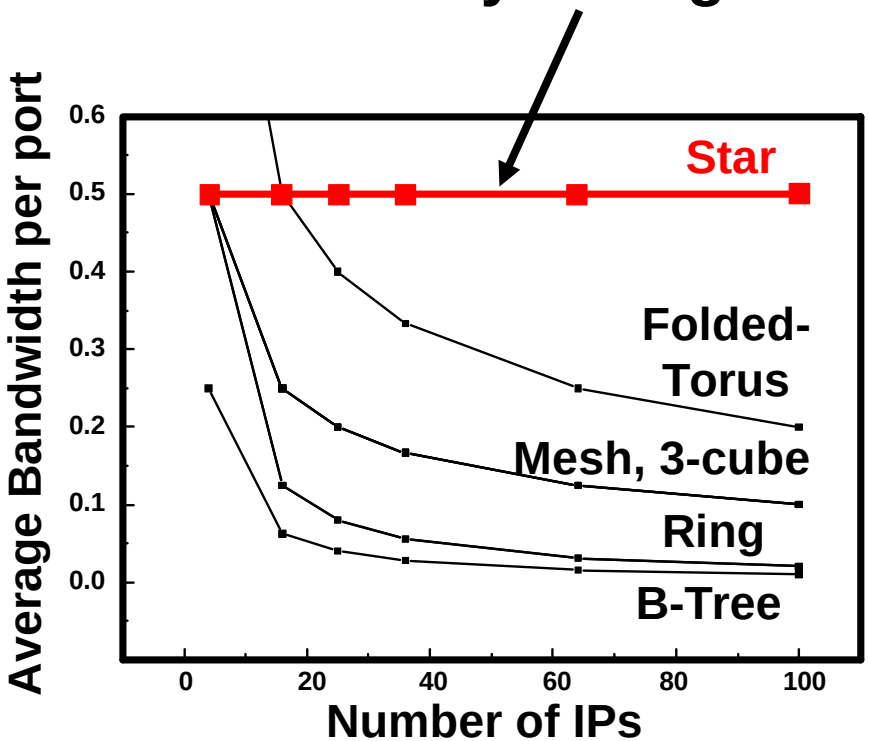
- **Techniques for**
  - **1) Bandwidth guarantee**
    - **Star-topology**
  - **2) Latency minimization**
    - **Source routing, NADR modification**
  - **3) Area reduction**
    - **Network partitioning, Up/Down Sampling**
- **Special features**
  - **4) Globally asynchronous system**
    - **Plesiochronous communication**
  - **5) Chip-chip packet transaction**
    - **Off-chip gateway**

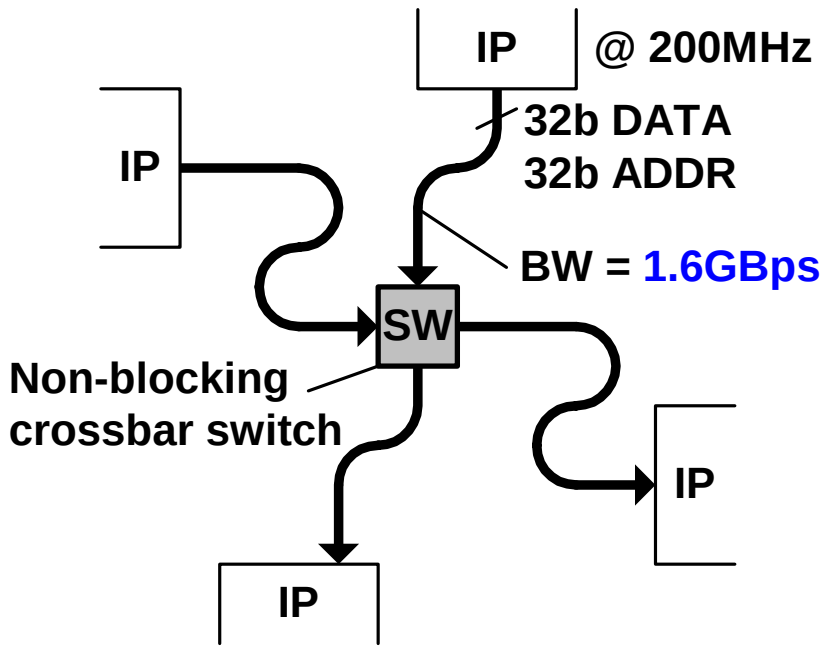
Overall Architecture of OCN



Bandwidth Guarantee: Star-topology

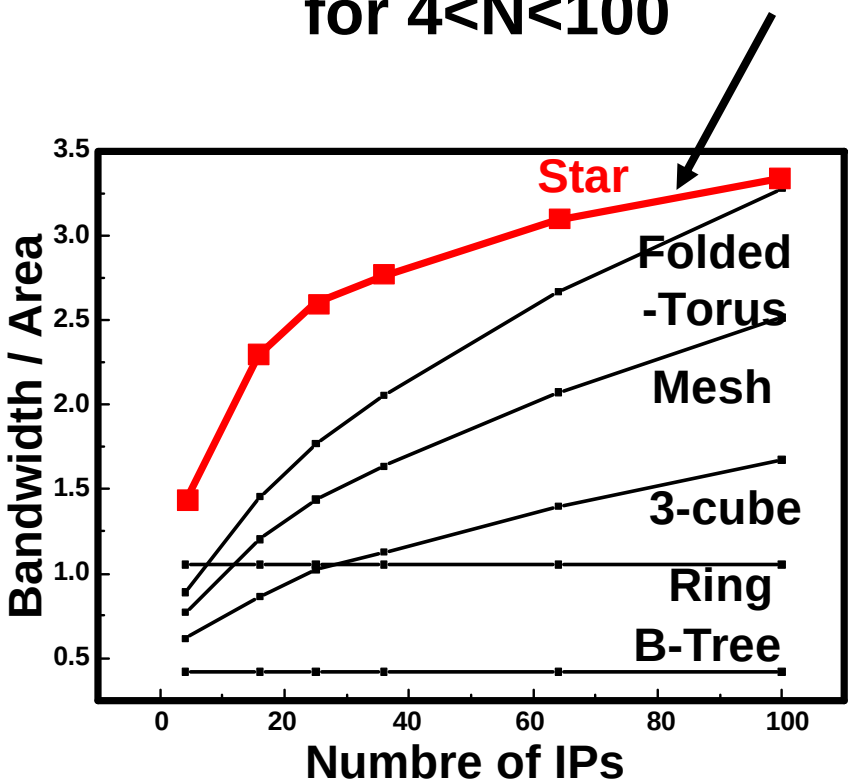
- **Sustains Full 1.6GBps / port bandwidth for any configuration**





Star-topology (Cont'd)

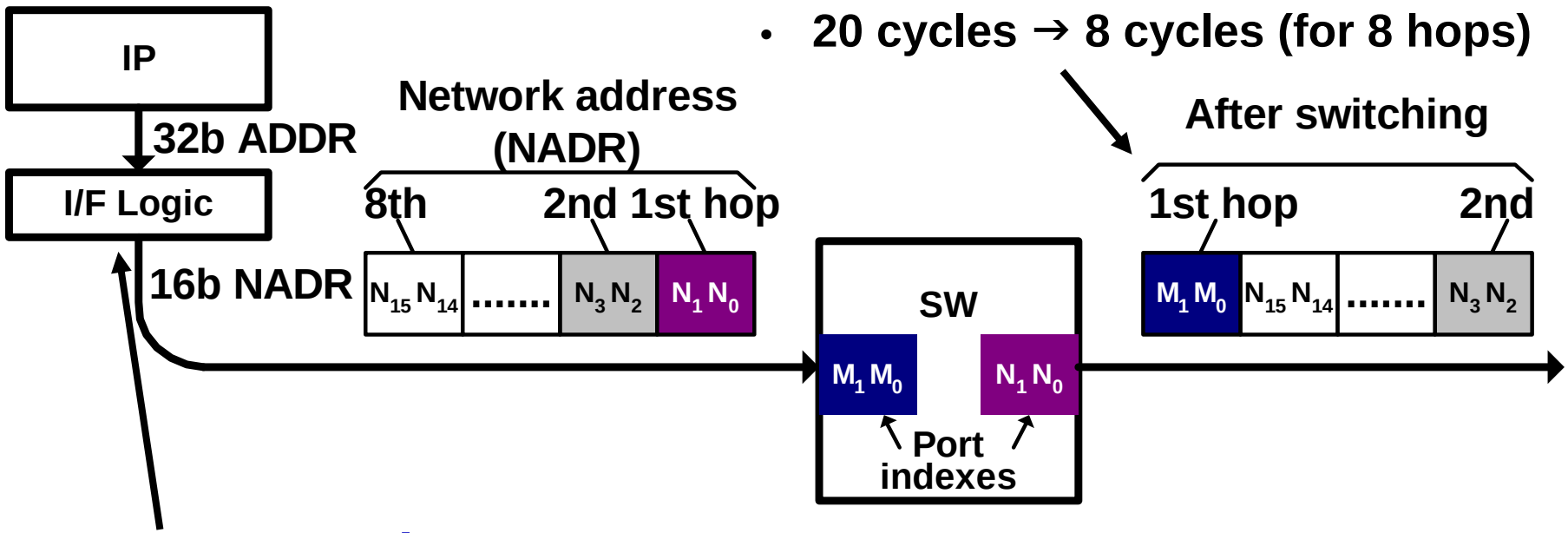
- **Shows the Best Performance/Cost metric for 4<N<100**



- **Network cost**
  - **Link + Switch area**
- **Evaluation environment**
  - **Square chip area**
  - **Uniformly distributed IPs**
  - **Orthogonal metal routing**
  - **Crossbar switch fabric**

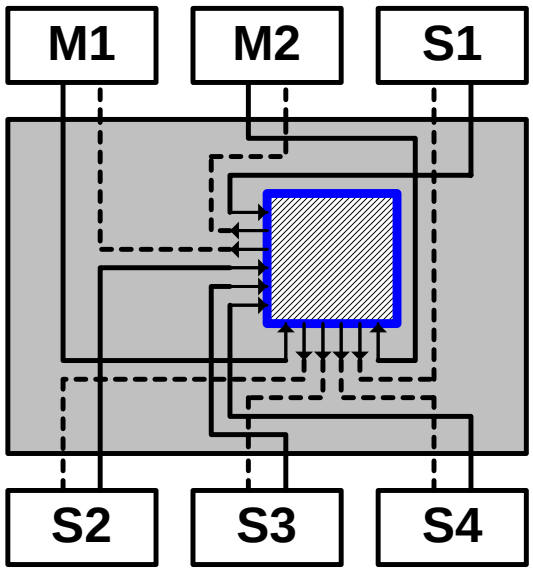
Latency Minimization: Routing Scheme

- **NADR modification**
  - **Fast cut-through switching**
  - **20 cycles → 8 cycles (for 8 hops)**

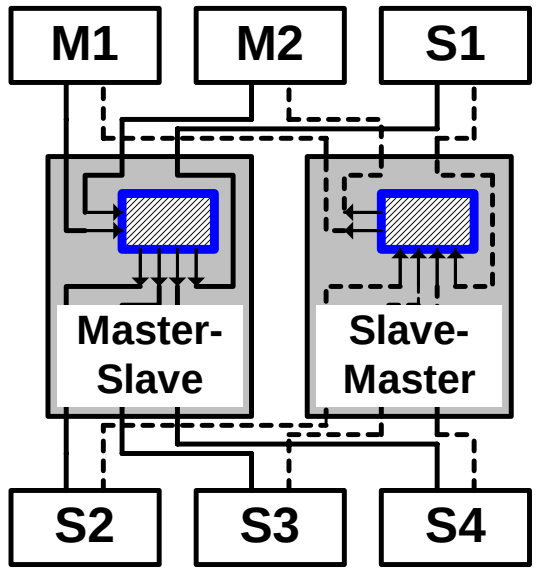


- **Source routing**
  - **Only 1 “Address Lookup” operation**
  - **Switches free from routing algorithm**

Area Reduction 1: Network Partitioning



Partitioning



$(m+n) \times (m+n)$
$= m^2 + n^2 + 2 \times m \times n$

Switch area

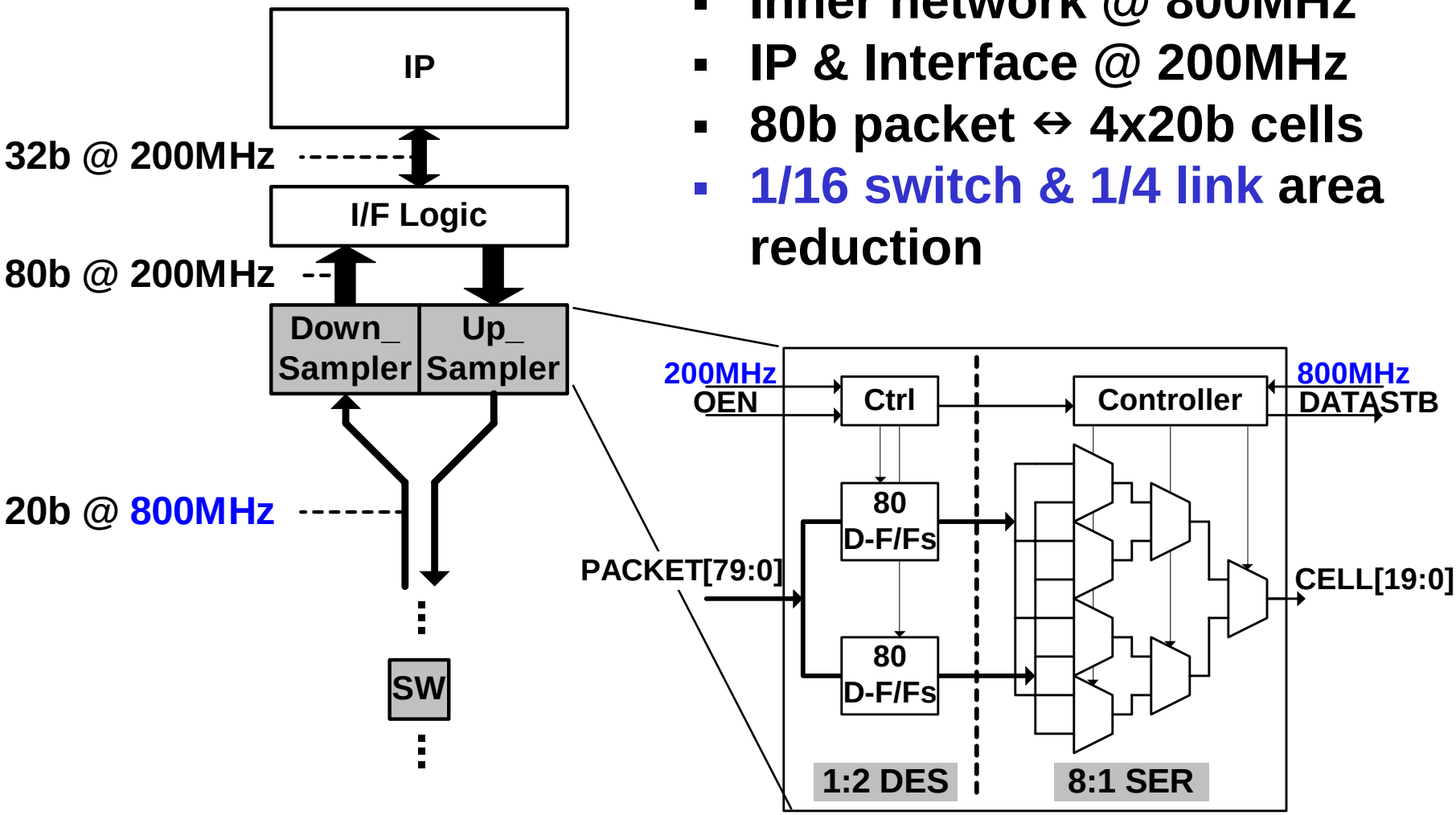
$2 \times (m \times n)$

m = # of Masters
n = # of Slaves

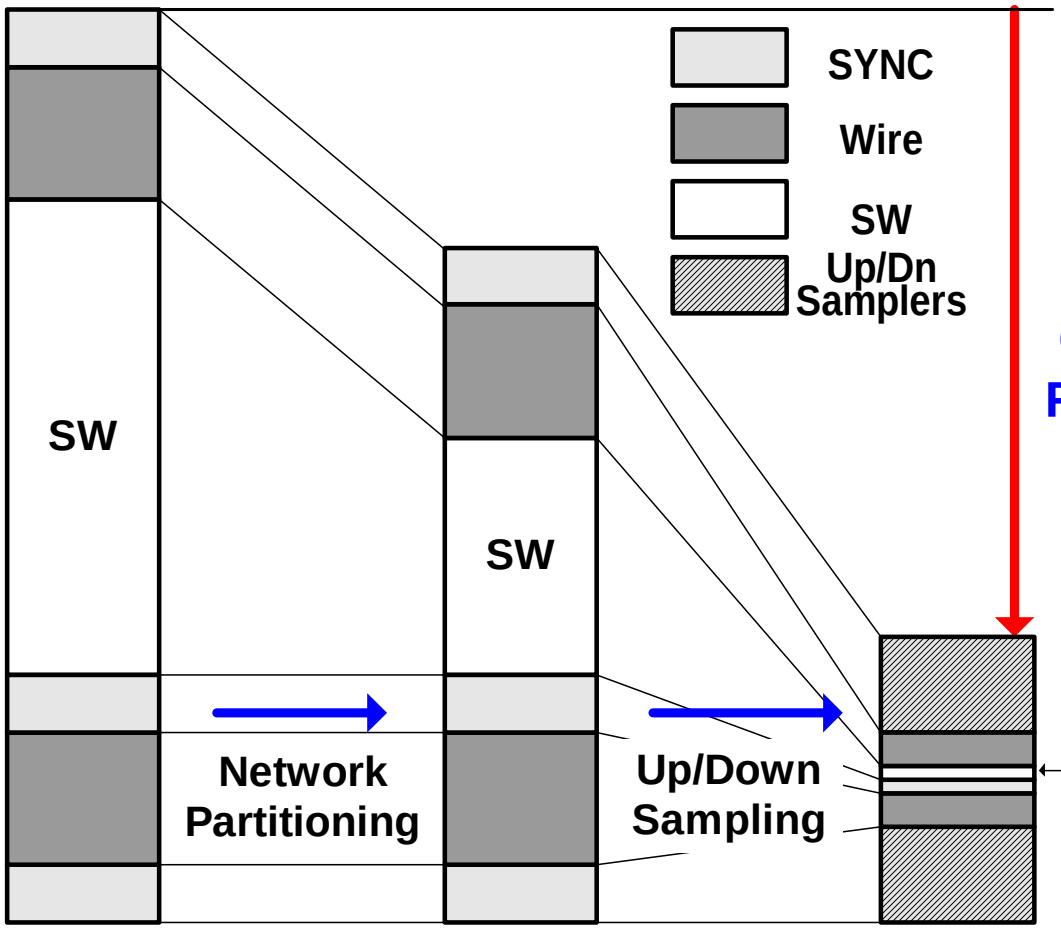
- **50% reduction of overall switch area (when m=2, n=2)**

Area Reduction 2: Up/Down Sampling

- **Inner network @ 800MHz**
- **IP & Interface @ 200MHz**
- **80b packet ⇔ 4x20b cells**
- **1/16 switch & 1/4 link area reduction**



Overall Area Reduction



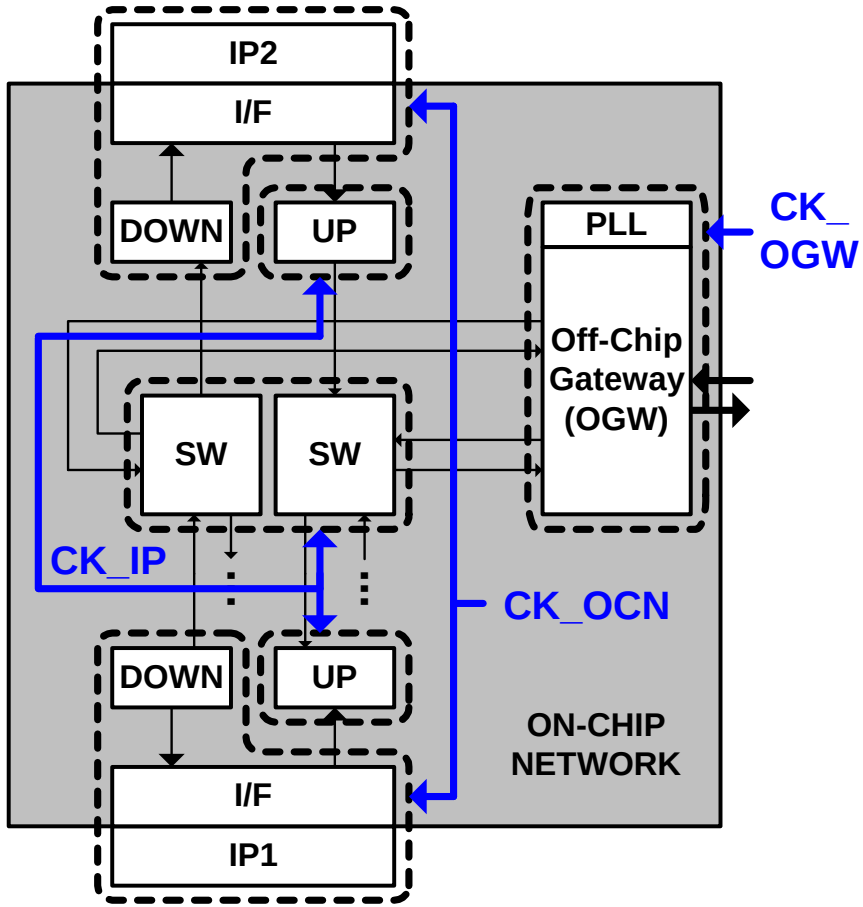
- **Switch area = $O(N^2)$**
- **Others' area = O(N)**

68% Area Reduction

- **78% area reduction for m,n=4**
- **85% for m,n=8**

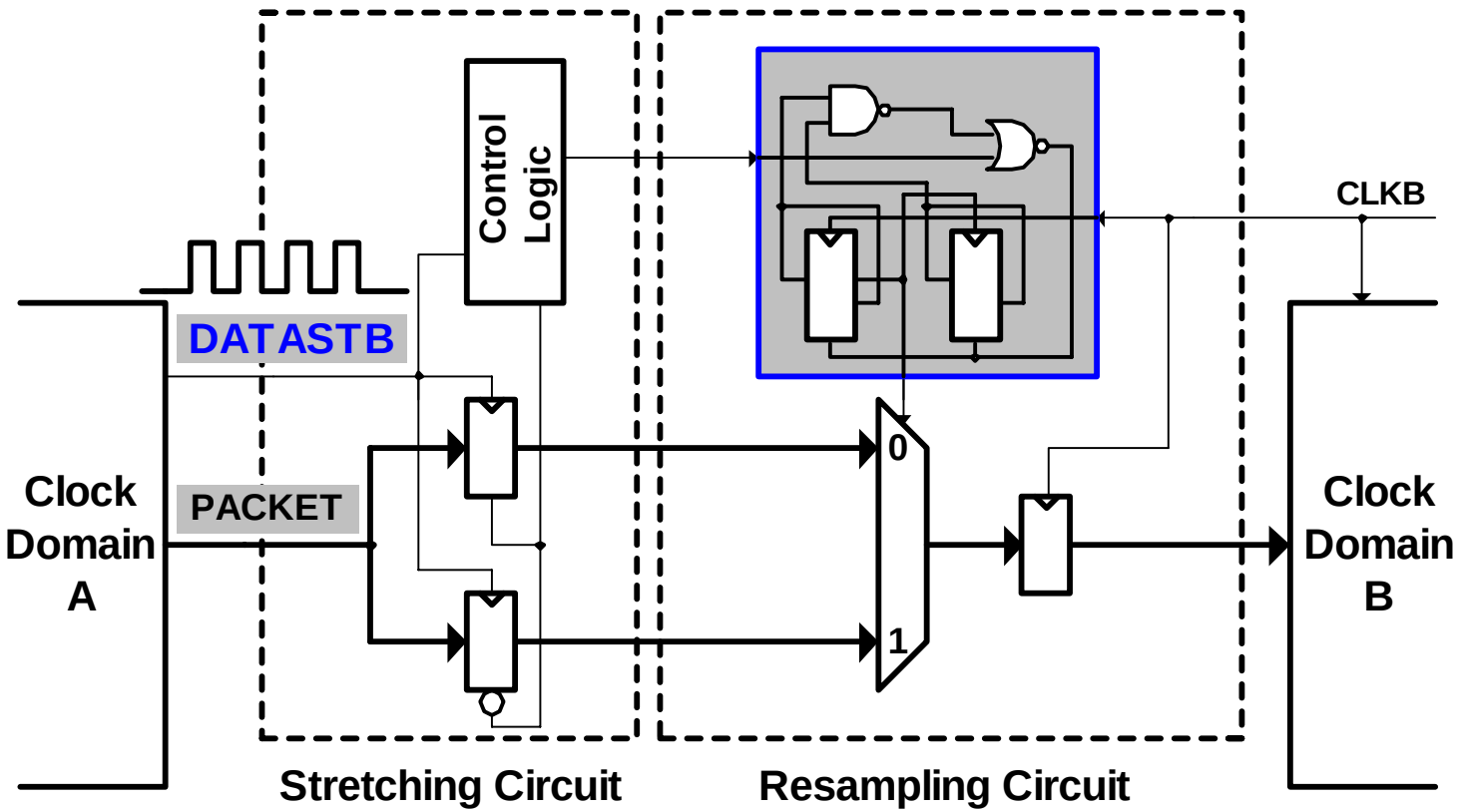
Special Feature 1: Plesiochronous Communication

- **Design environment**
  - **Global synchronization ignored**
  - **Multiple clock sources**
- **Clock domain**
  - **Region located physically apart from others**
  - **Region with different clock source**



Plesiochronous Communication (Cont'd)

- **Plesiochronous packet transaction**
  - **Synchronizer b/w different clock domains**
  - **DATASTB for data recovery**



Special Feature 2: Chip-chip Communication

(1) Off-chip gateway

→ **Deals with chip-chip packet transaction**

(2) Packet transaction over Bus

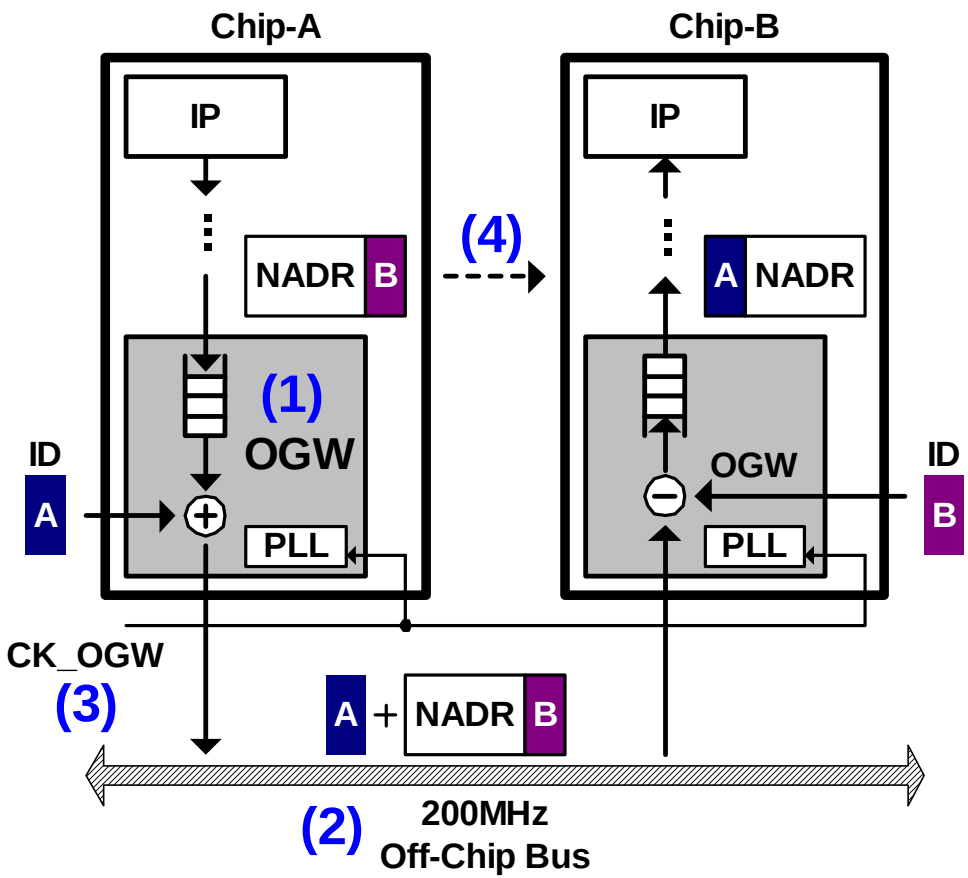
→ **No additional off-chip component**

(3) Synchronized operation

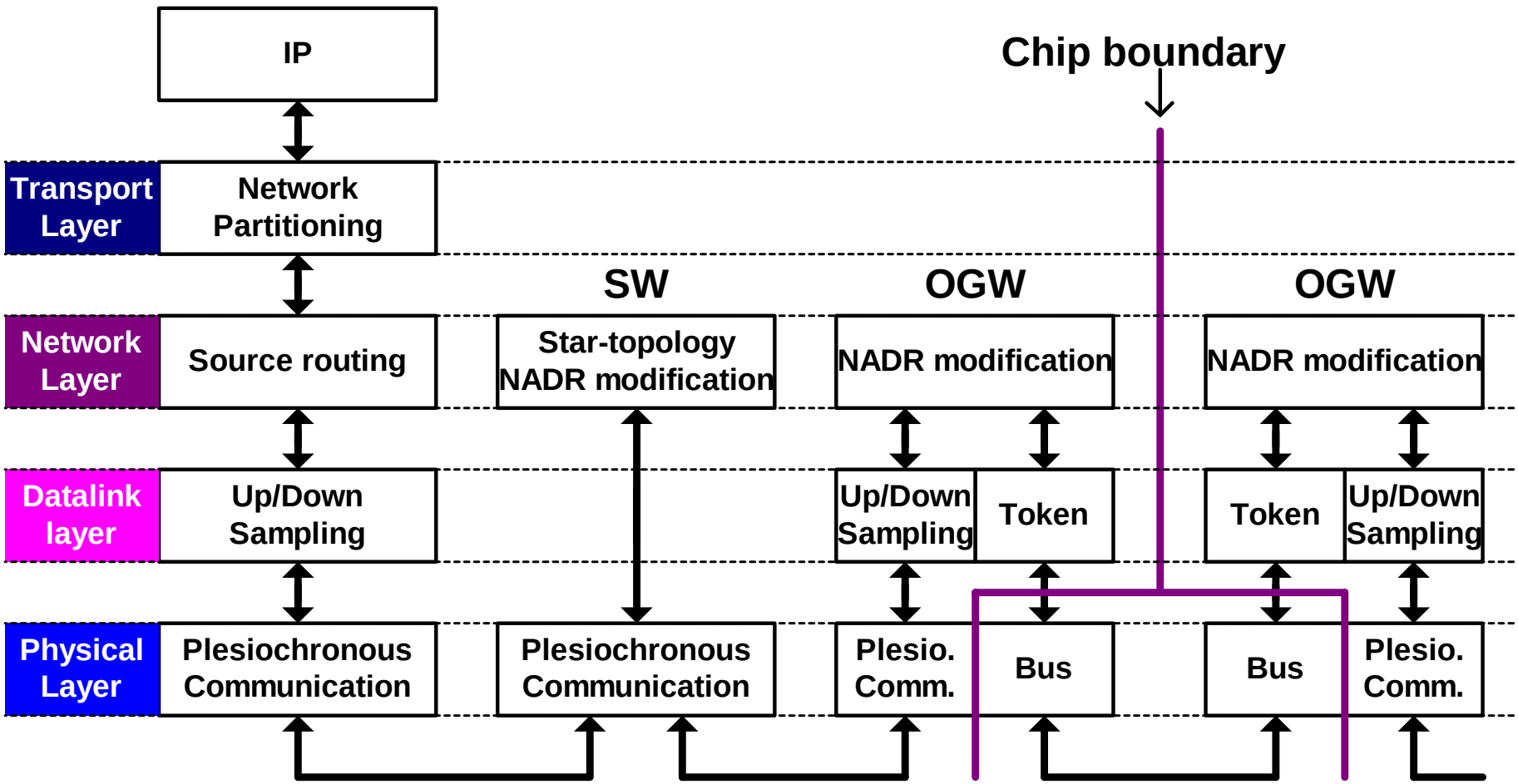
→ **Predictable operation**

(4) Consistent NADR convention

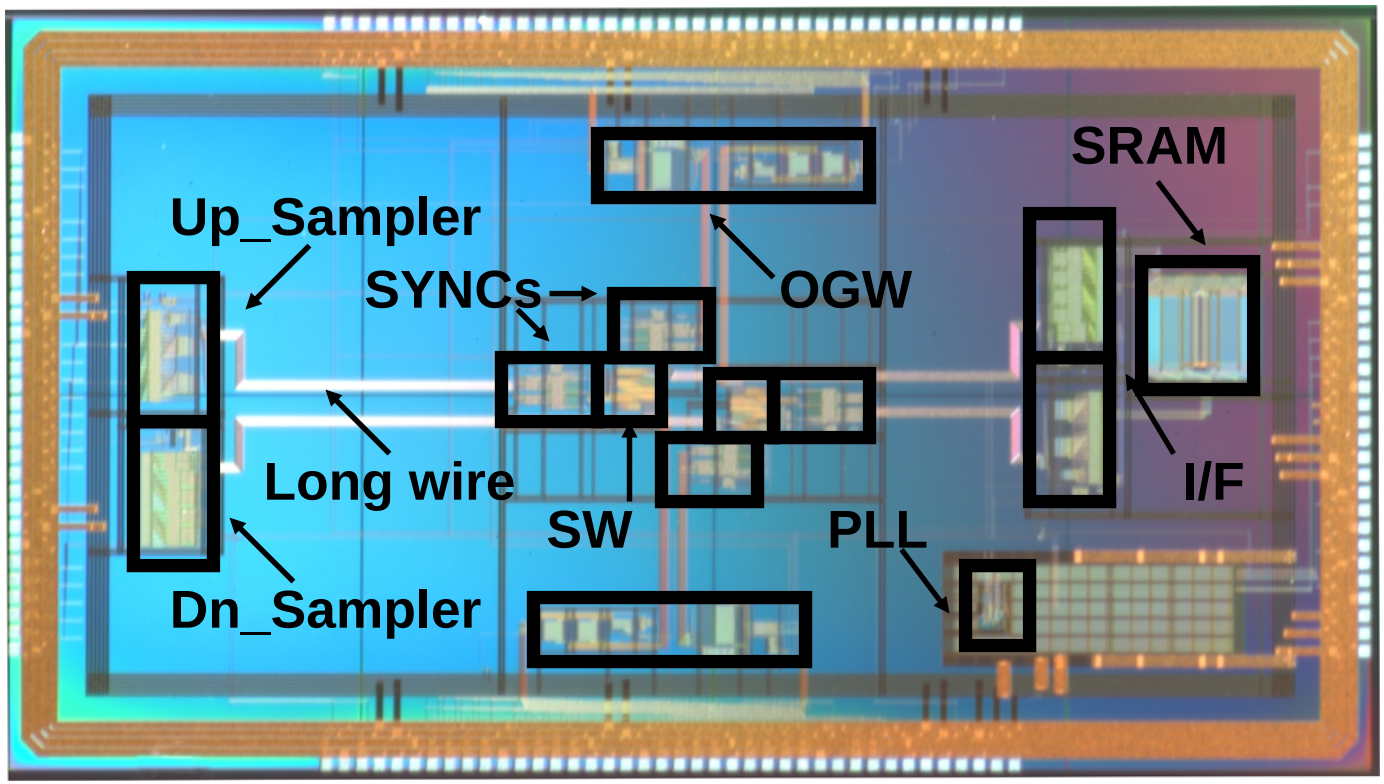
→ **Seamless off-chip packet transaction**



Feature summary

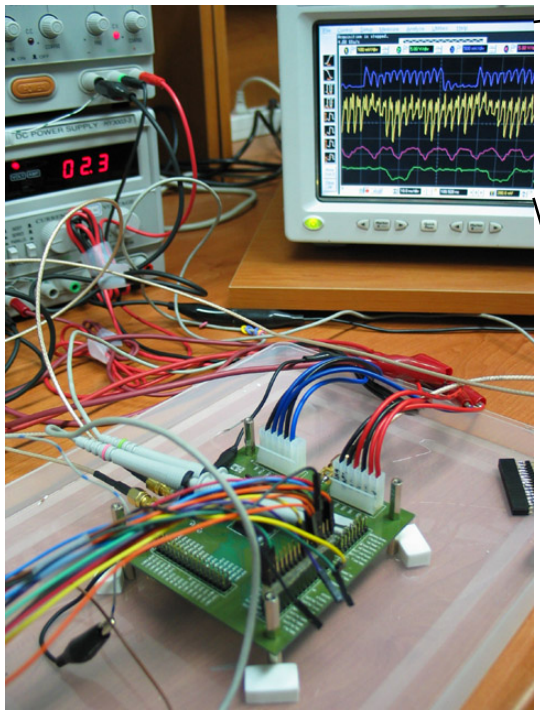


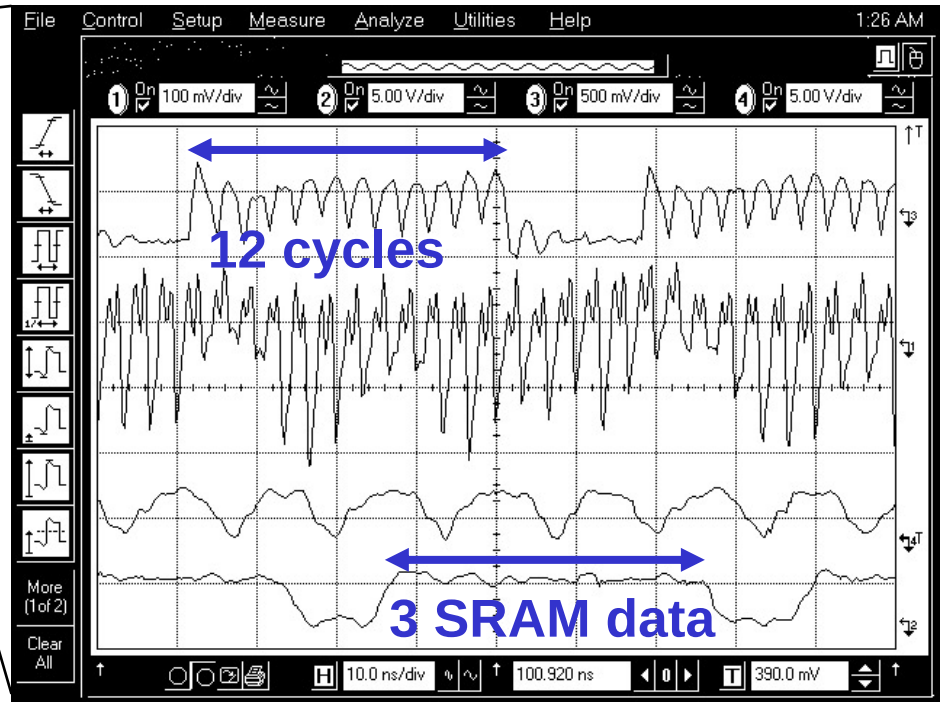
Die Photo



Die size	10.8 x 6.0 mm^2
Technology	0.16um DRAM technology with 3 Al
Transistor count	81,000 (without 1kB SRAM)
Power (OCN)	264mW @ 2.3V

Measurement Results: Single chip



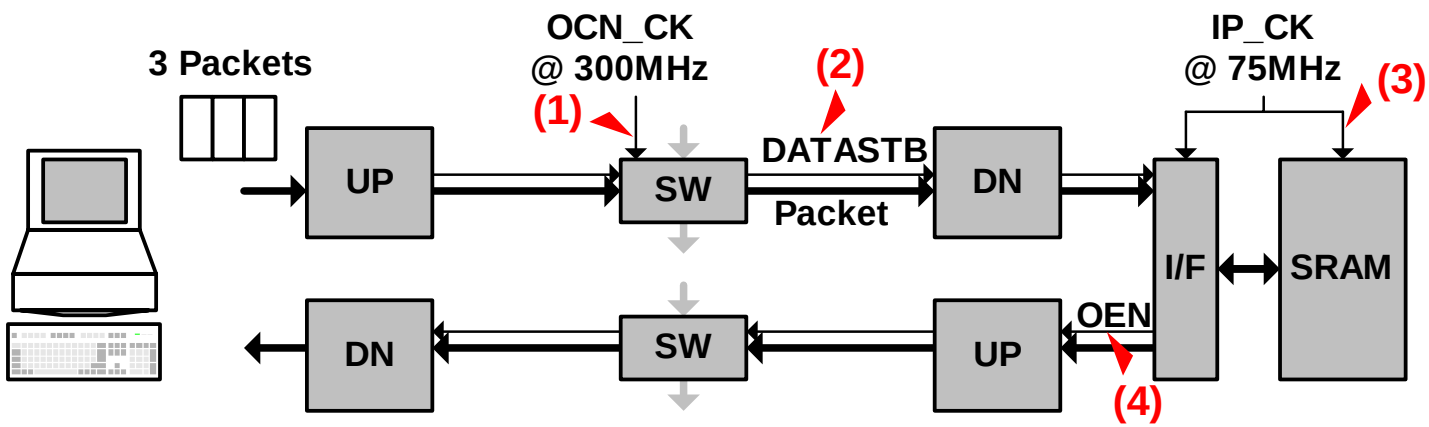


(2)

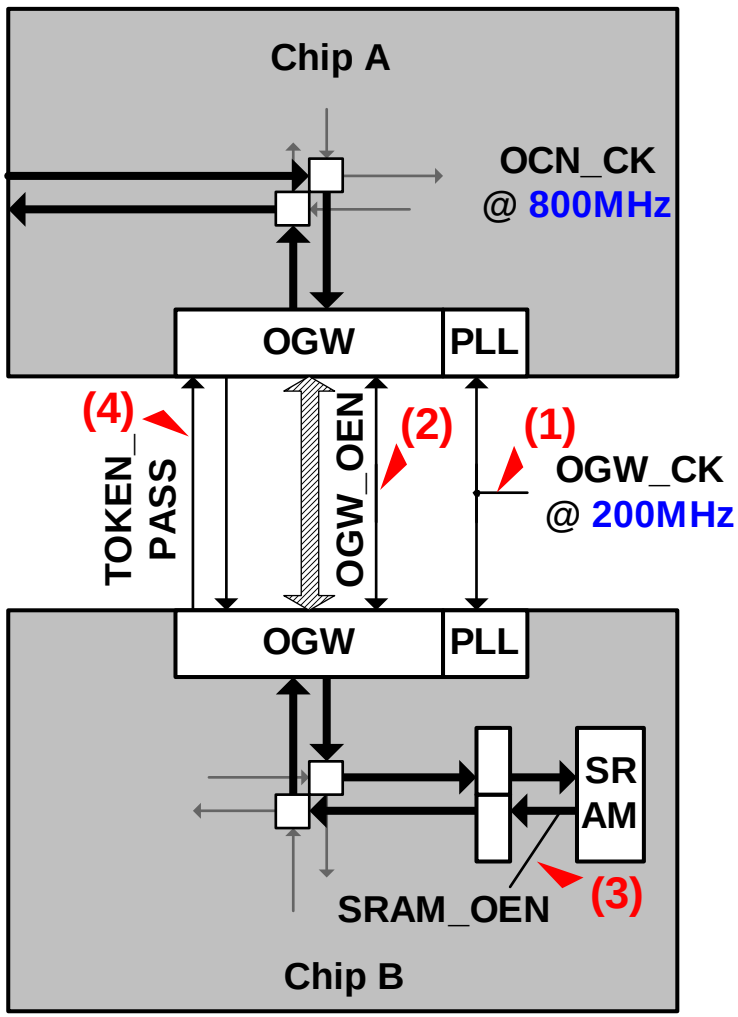
(1)

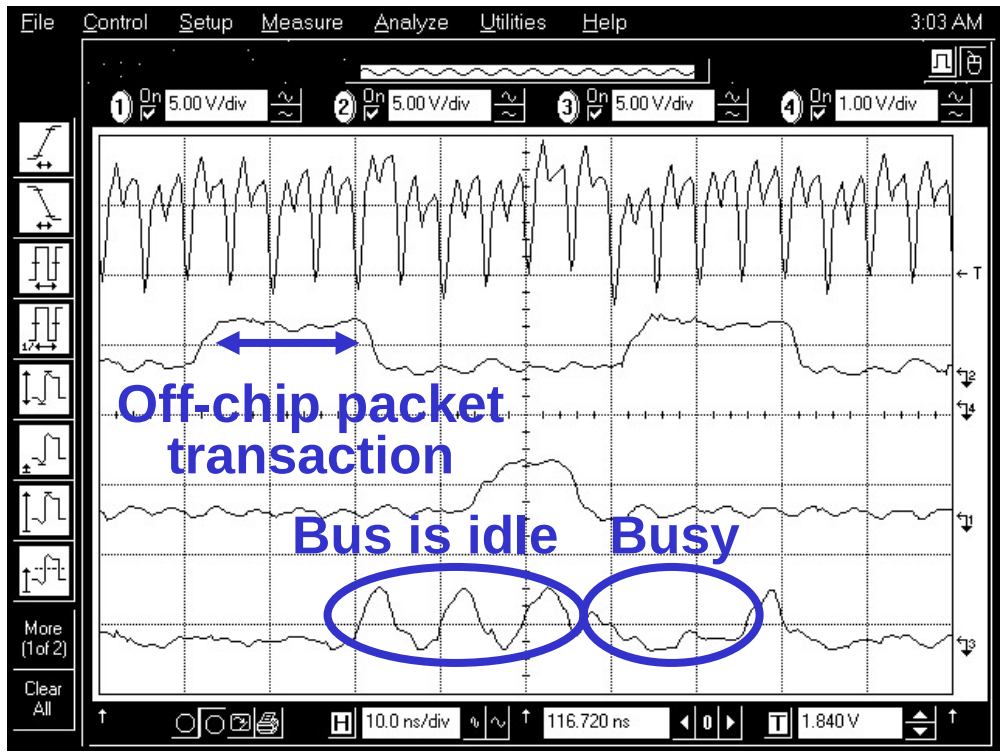
(3)

(4)



Measurement Results: Chip-chip

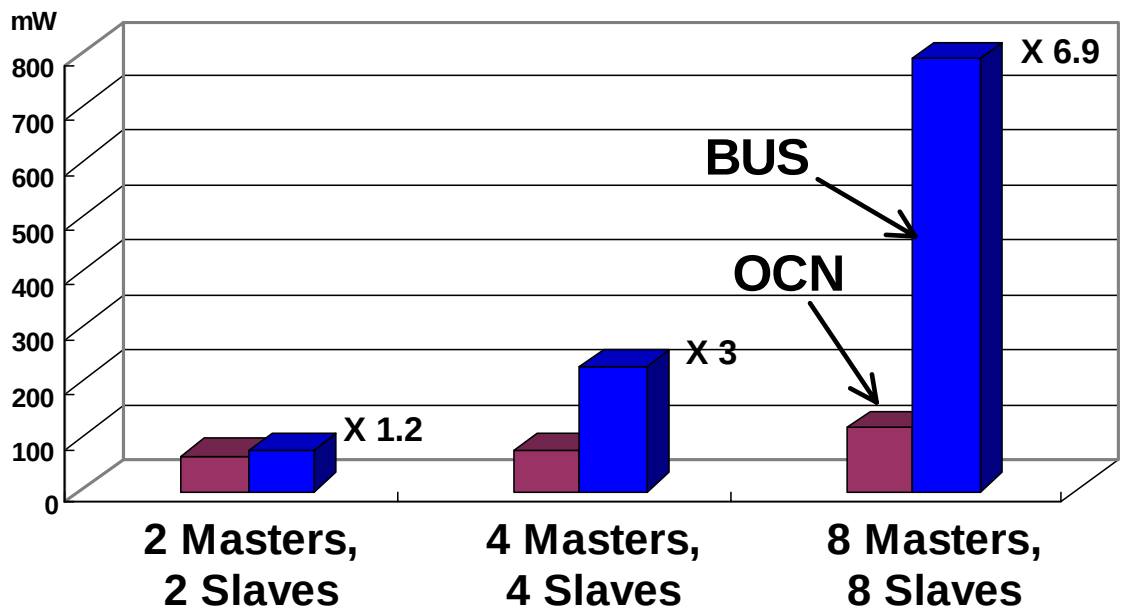




Power comparison with Bus

- $BW_{BUS}(f_{BUS}) = BW_{OCN}(f_{OCN})$
- N = Min {# of Masters, # of Slaves}
- $P_{OCN} = \alpha \cdot f_{OCN} \cdot (C_S + C_D + C_{OVERHEAD}) \cdot V^2$
- $P_{BUS} = \alpha \cdot (N \cdot f_{OCN}) \cdot (C_S + C_{D1} + C_{D2} + \ldots + C_{DN}) \cdot V^2$

Power consumption for 1 packet transaction



Conclusion

- An 800MHz star-connected on-chip network
  - Provides 1.6GB/s per port bandwidth
  - Network partitioning and Up/Down sampling saves 68% overall network area
  - Utilizes efficient routing scheme
  - Plesiochronous communication for GALS system
  - OGW for seamless chip-to-chip communication