Asynchronous Three-Dimensional Networks-on-Chip

Abbas Sheibanyrad

Frédéric Pétrot

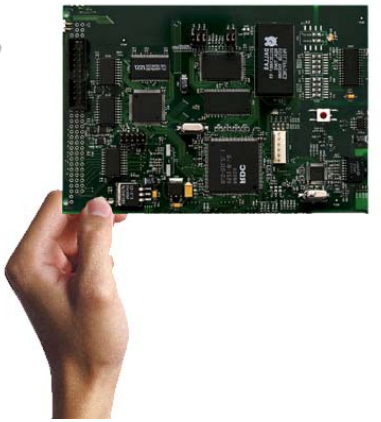
Outline

- Three-Dimensional Integration
- Clock Distribution and GALS Paradigm
- Contribution of the Third Dimension
- Asynchronous 3D-NoC
- An Attractive Idea !
- Conclusion

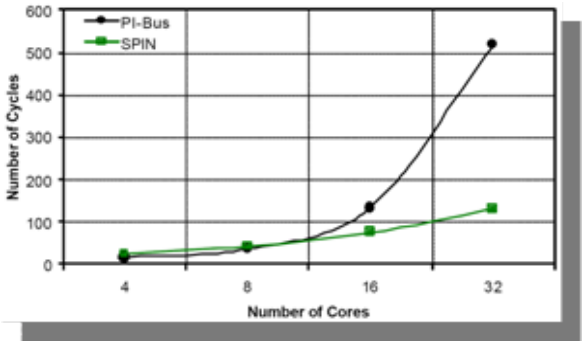
Technology Evolution !

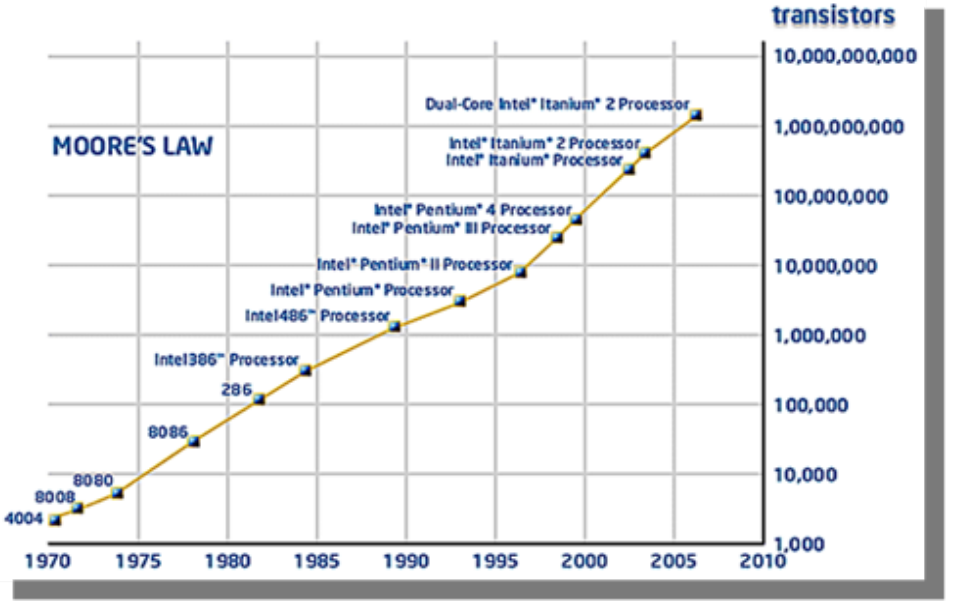
- Evolution of the fabrication technology
  - Integration of systems with billions of transistors in only one chip
  - Hundreds and even thousands of components
  - Key role of the communication infrastructure
  - Scalability

System-on-Board



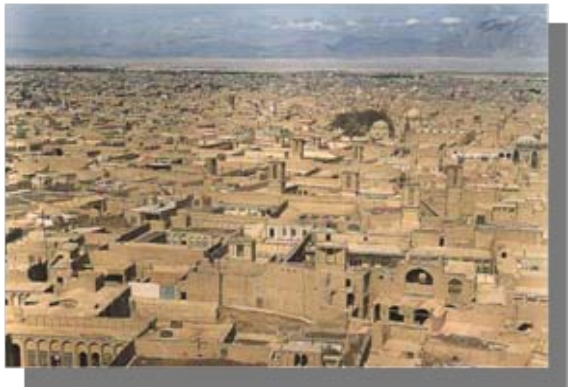
System-on-Chip



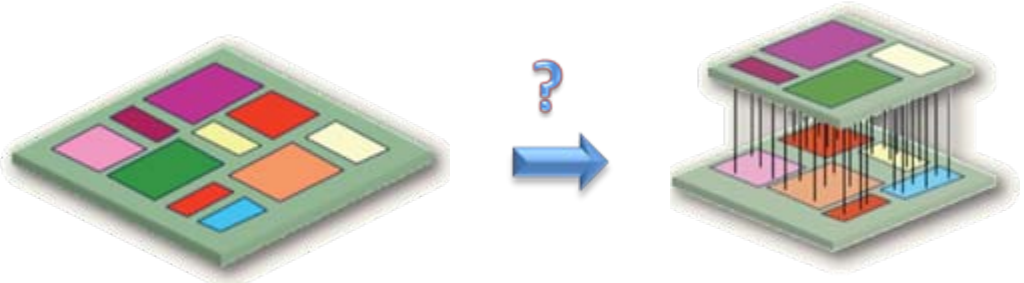


… but, the land becomes expensive !

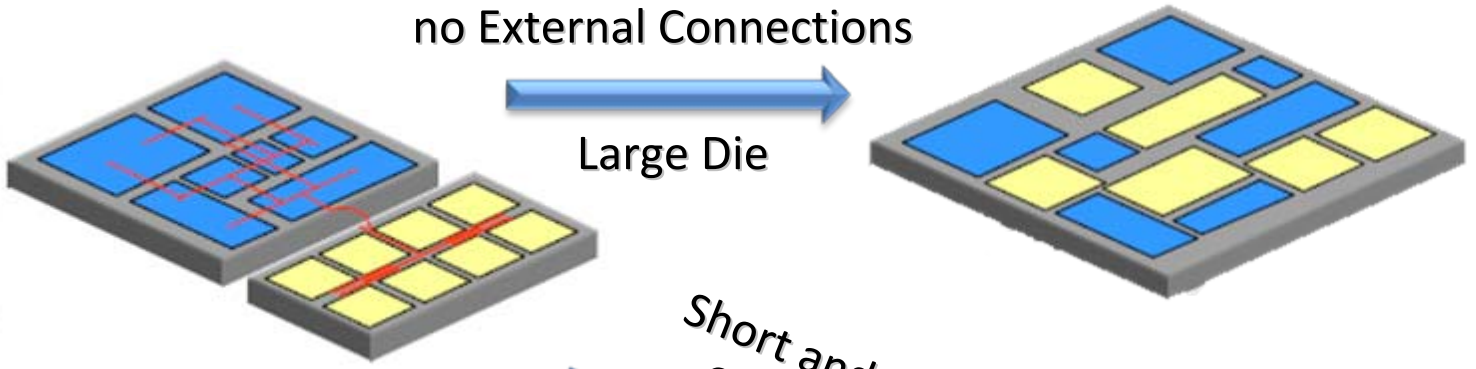
- As the land becomes more and more expensive, there is a trend to build vertically rather than horizontally
  - Increase the density
  - Decrease the length and the number of long paths

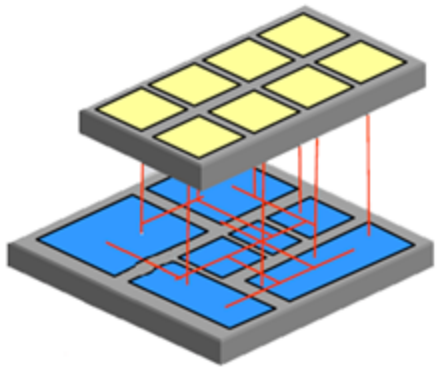


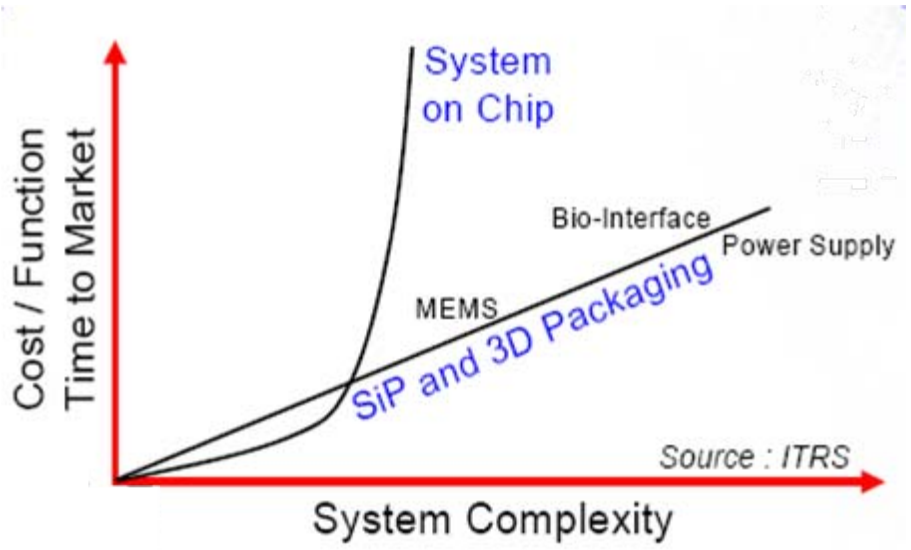
- Any Limitations on the third dimension ?
  - Technological Constraints



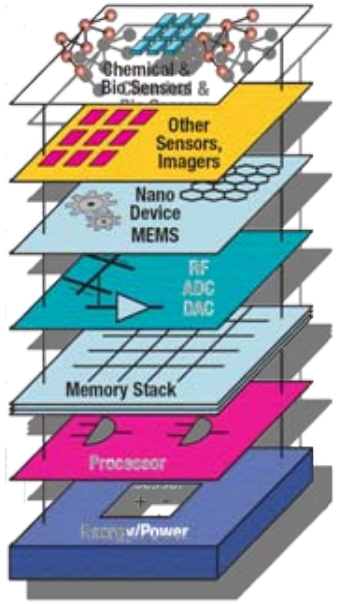
Why not 3D Integration of Silicon Dies ?





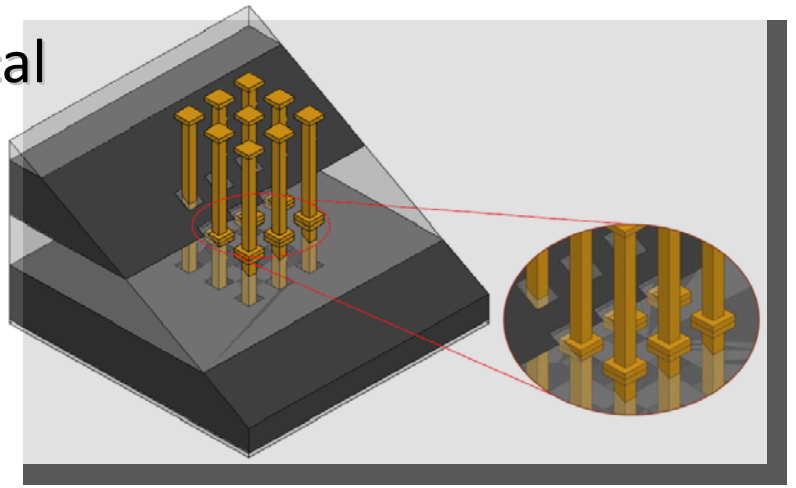


Possibility of Integrating Different Dies with Different Technologies in a same Package



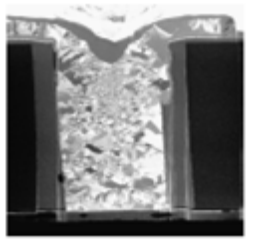
Through-Silicon-Via

- The most promising Technology of Vertical Interconnection
  - Low Resistance and Capacitance
  - High Bandwidth
  - Low Power Consumption

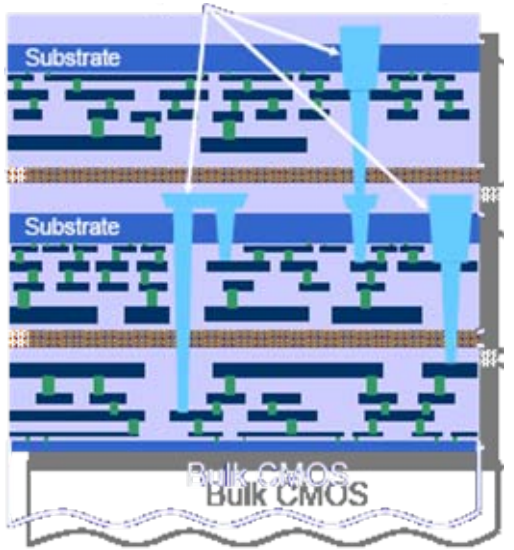


- Via-First (higher density of TSVs)
  - Diameter ≈ 5 µm (*IMEC)
  - Pitch ≈ 10 µm
  - Depth ≈ 20-50 µm

- Via-Last (lower cost of the process)
  - Diameter ≈ 35 µm
  - Pitch ≈ 60 µm
  - Depth ≈ 40-150 µm



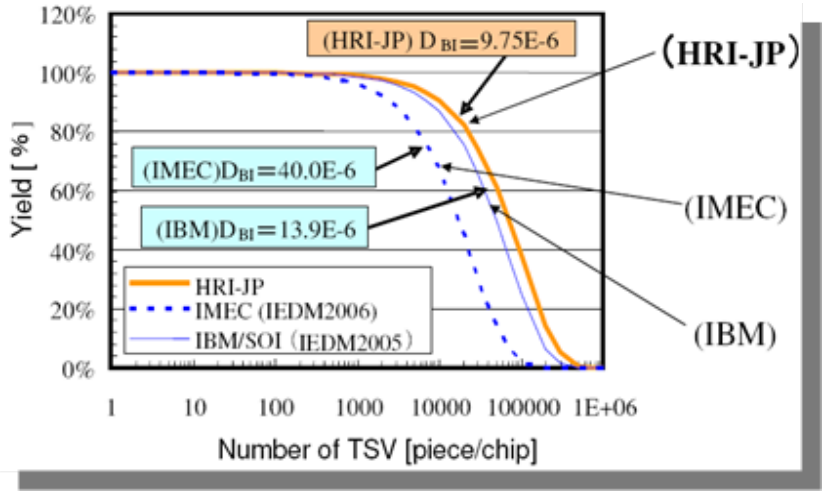
TSVs



… but, what is the Reliability of TSVs ?

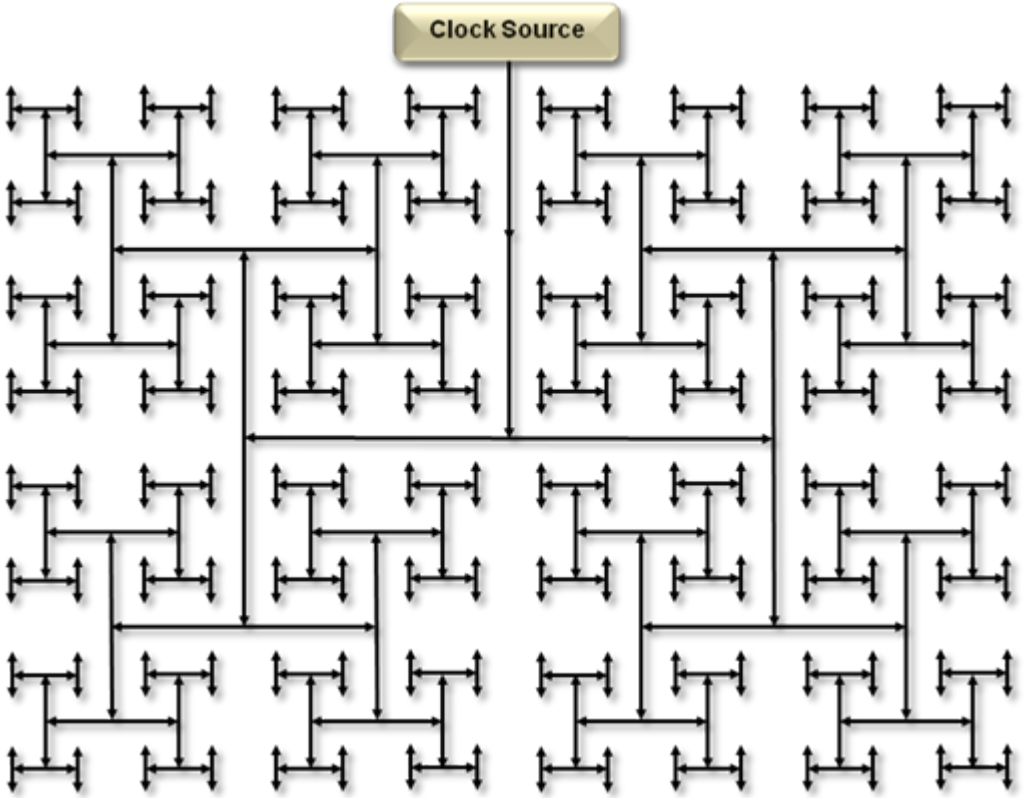
- Important risk of failure due to several additional fabrication steps (a potential reduction on the Yield)
  - Misalignment
  - Dislocation
  - Void formation
  - Oxide film formation over Copper interfaces
  - Pad detaching
  - Defects due to temperature
  - …

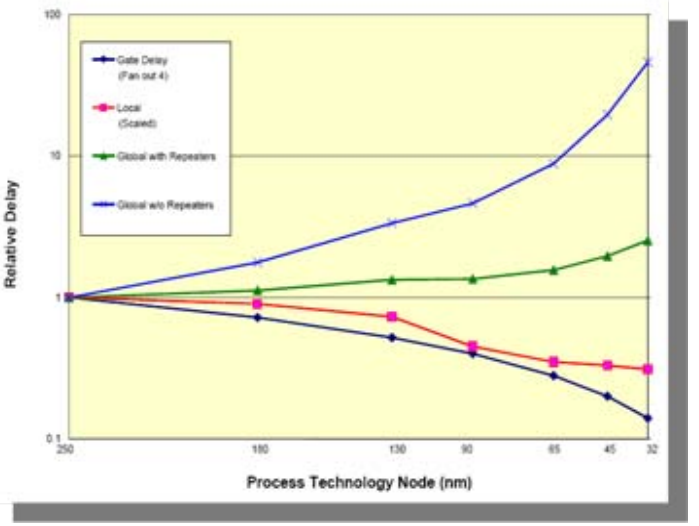
The Three-Dimensional Integrated Circuits are limited by the number of TSVs to be exploited



Clock Distribution

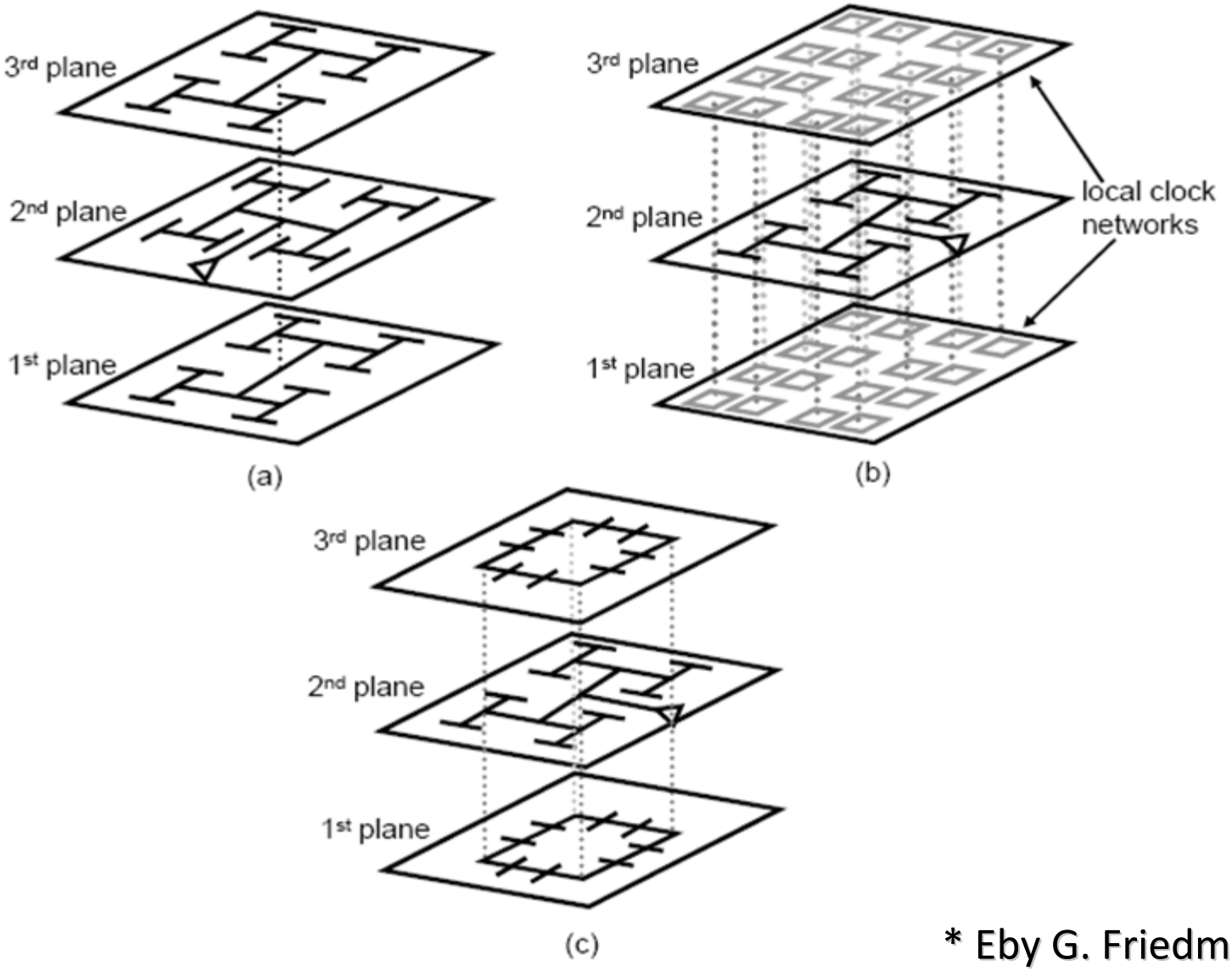
- Deep Submicron Technologies
  - Aggravation of physical problems
  - Predominant effect of long wires on delay and consumption





- Nightmare of Global Synchronization
  - Impossible Global Distribution of a single clock signal over a chip
  - Fabrication Process Variation
  - Temperature Variation

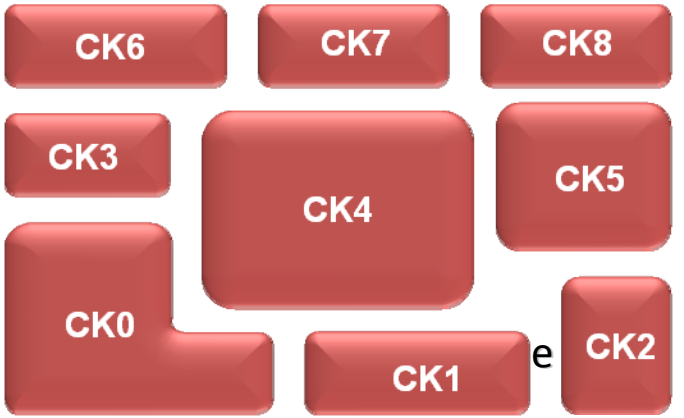
Clock Distribution in 3 Dimensions

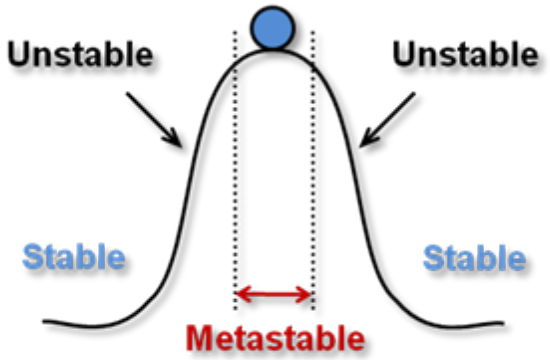


* Eby G. Friedman

GALS always demanded !

- The GALS paradigm (Globally Asynchronous Locally Synchronous) is an attractive solution
  - Several domains clocked independently
- Networks-on-Chip are the most Structured Approaches
  - The Network is the asynchronous global part of the system
  - The subsystems are the synchronous local parts system



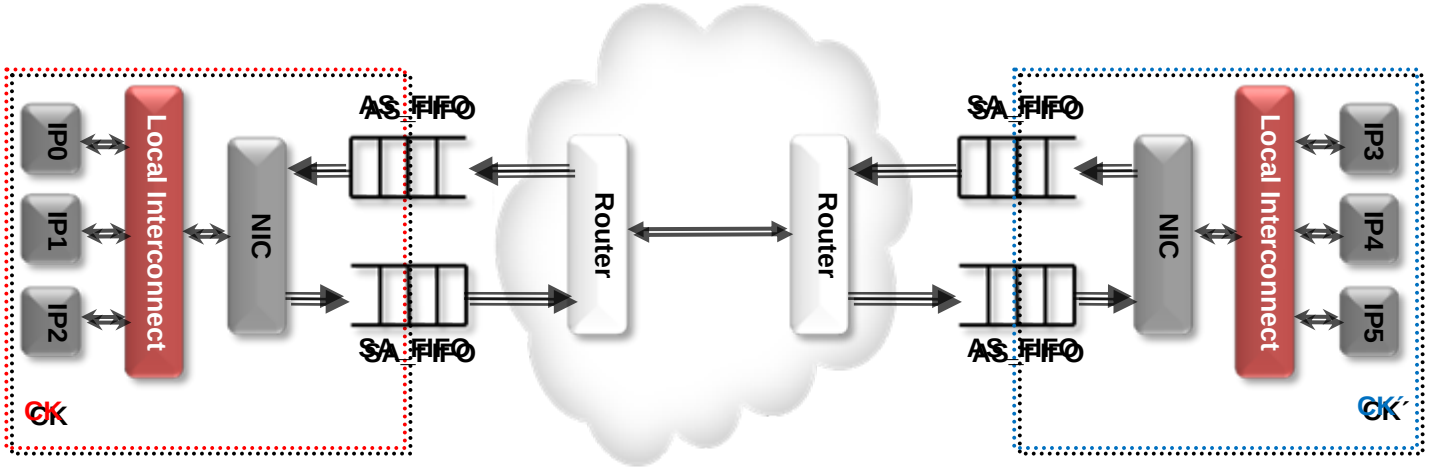


… but, how can two separately clocked domains communicate in a reliable manner ?

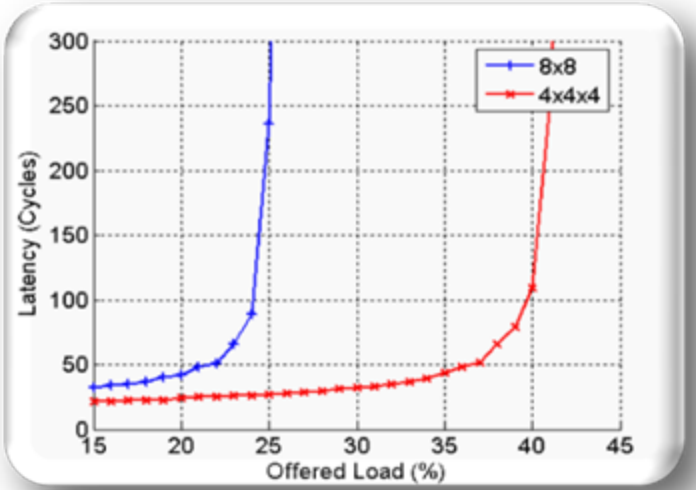
- Metastability, an unavoidable state of a bistable system, is the major problem of the GALS architectures

ASPIN: a fully Asynchronous NoC

- The need of synchronization reduced to the network interfaces
  - Special FIFOs: Async-to-Sync and Sync-to-Async
  - An End-to-End latency much lower than the multi-synchronous version
- As fast as possible and independent from the rest of the circuit
  - Saturation threshold improved compared with the multi-synchronous version
- The almost zero dynamic Power Consumption in the idle state
- Scalability and Reusability in a Plug & Play fashion and independent from the size of subsystems



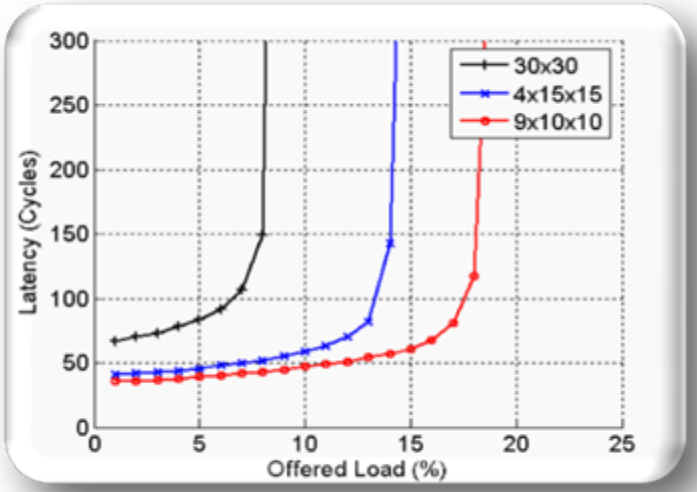
Contribution of the Third Dimension

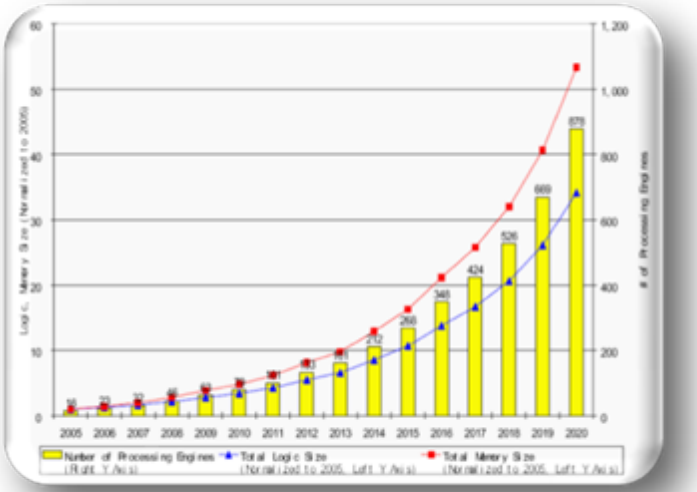


	Number of Nodes	Switch Degree	Network Diameter	Number of Channels	Number of Vertical Channels	Number of Bisection Channels	Load of the Busiest Channels [1]
2D-Mesh	$N = n^2$	5	$2\sqrt{N}$	$6N - 4\sqrt{N}$	0	$2\sqrt{N}$	$C \times \frac{1}{4}\sqrt{N}$
3D-Cube	$N = m^3$	7	$3\sqrt[3]{N}$	$8N - 6\sqrt[3]{N^2}$	$2N - 2\sqrt[3]{N^2}$	$2\sqrt[3]{N^2}$	$C \times \frac{1}{4}\sqrt[3]{N}$

[1] Assuming uniform destination distribution and dimension-ordered routing, C is the average load injected to the network by each node

Contribution of the Third Dimension



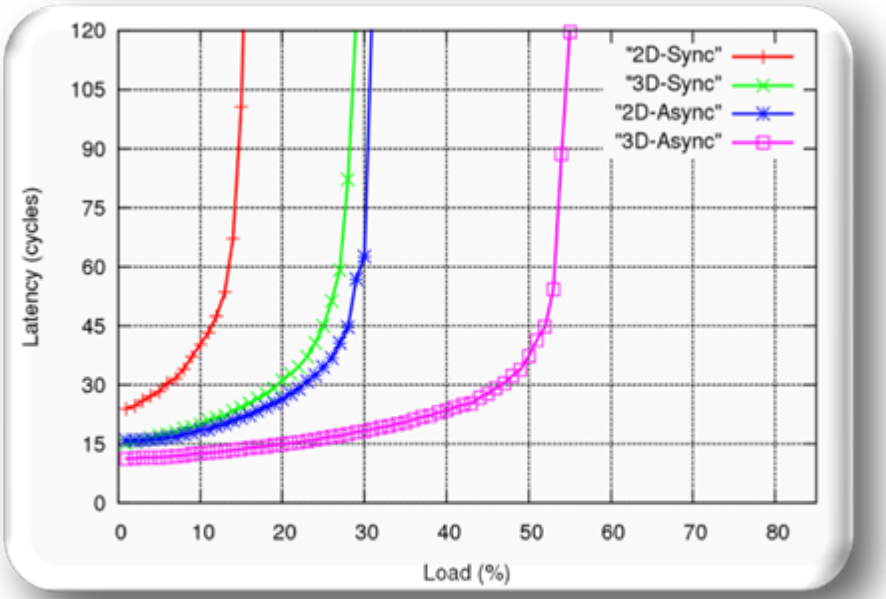


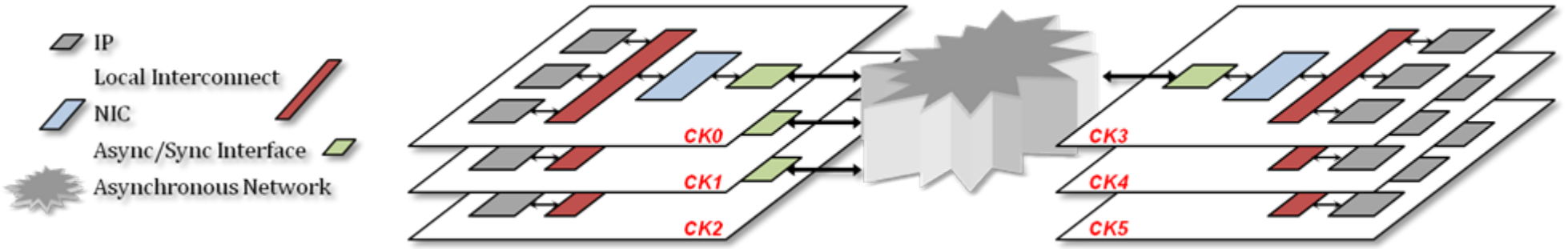
	Number of Nodes	Switch Degree	Network Diameter	Number of Channels	Number of Vertical Channels	Number of Bisection Channels	Load of the Busiest Channels [1]
30x30	900	5	60	5280	0	60	$C \times \frac{1}{4} \times 30$
4x15x15	900	7	34	6510	1350	120	$C \times \frac{1}{4} \times 15$
9x10x10	900	7	29	6640	1600	180	$C \times \frac{1}{4} \times 10$

[1] Assuming uniform destination distribution and dimension-ordered routing, C is the average load injected to the network by each node

Asynchronous 3D Network

- Insensitive to Delay Variation due to Temperature Variation or Process Variation
- Exploitation of the whole high Bandwidth of TSVs
- Speed ratio of 2 as a worst-case assumption
  - Using STMicroelectronics 90nm GPLVT transistors, 400MHz as the maximum clock frequency of usual SoCs
  - Using the same technology, 1100 Mflits/s as throughput of an asynchronous NoC

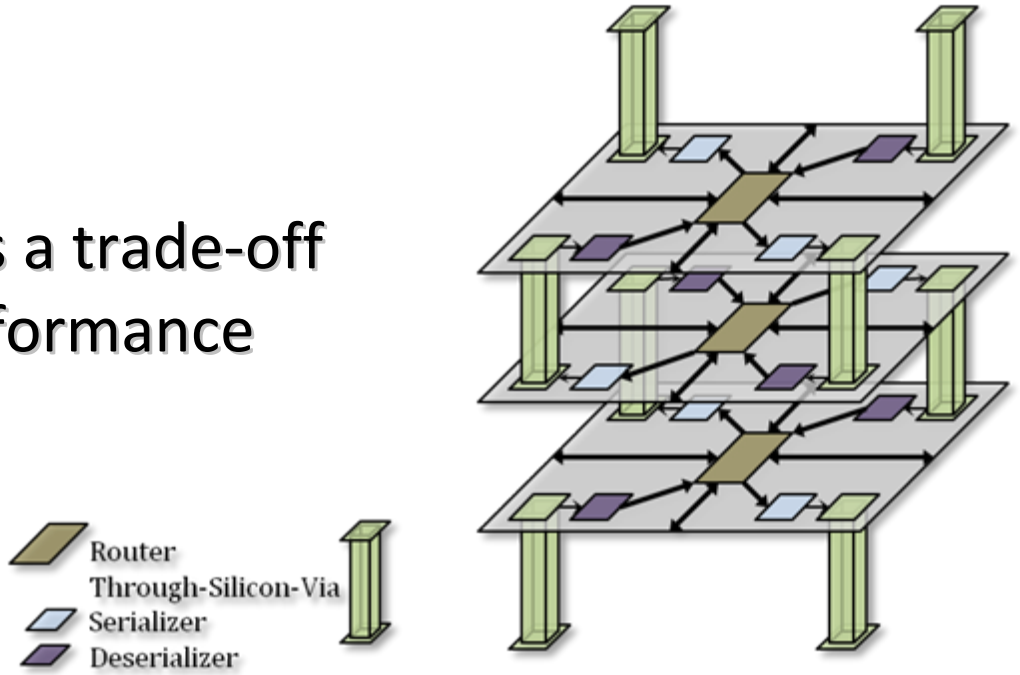




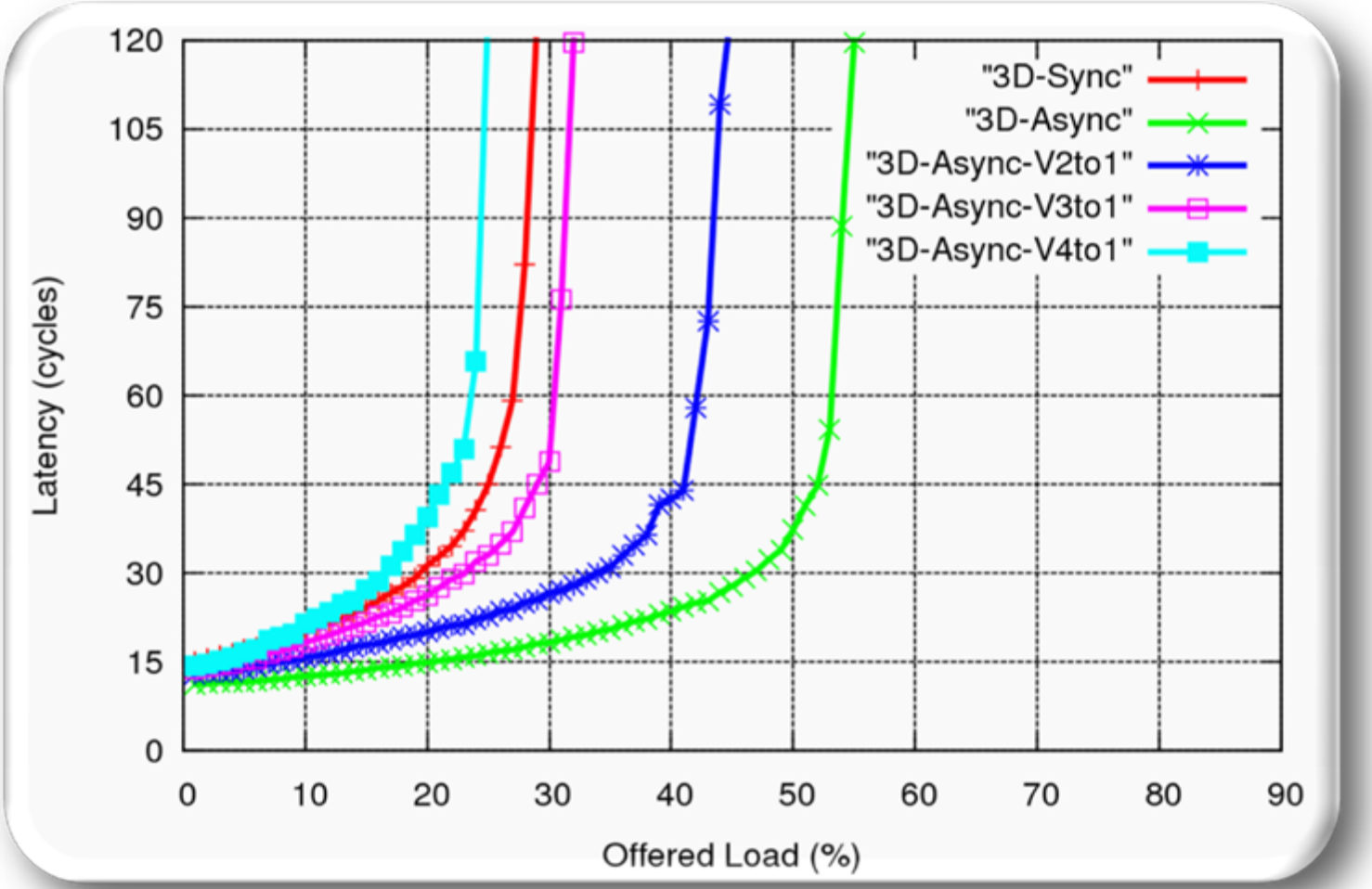
Why not Serialized Vertical Links ?

- Remembering
  - Using TSVs guarantees a faster vertical data transfer with lower power consumption than horizontal links in moderate size
  - but, the Pitch of TSVs is large, and, several additional steps of TSV fabrication add a potential reduction of the Yield
  - Only a small fraction of the capacity of vertical link is exploited

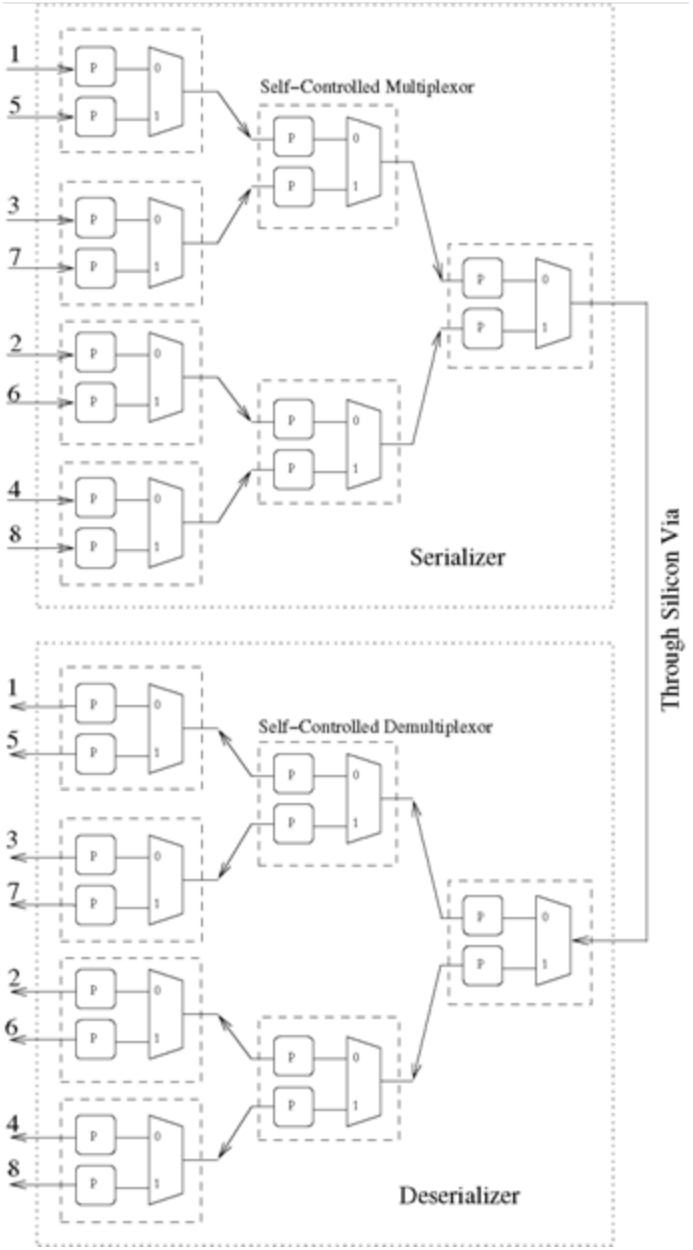
- Serialization of data on TSVs is a trade-off between the cost and the performance

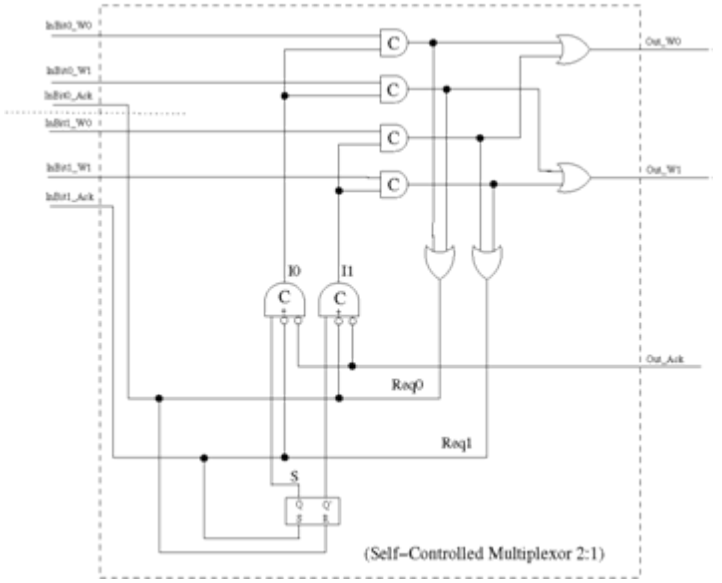


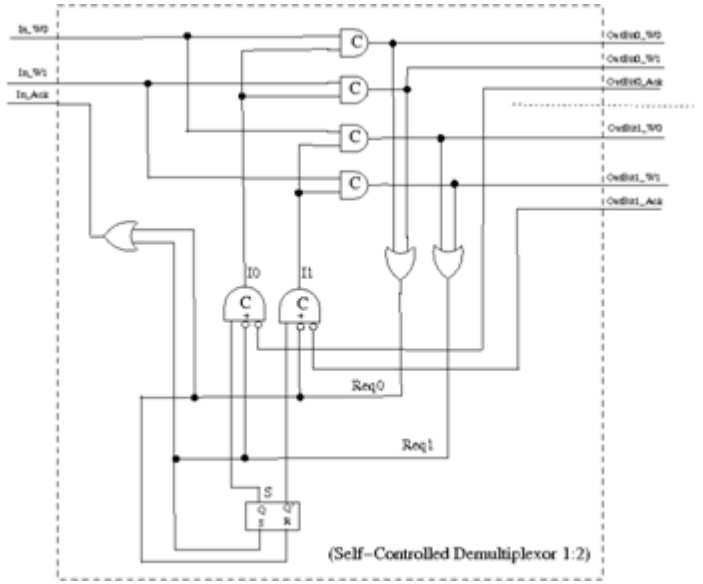
Vertically Serialized Asynchronous 3D-NoC



Circuit Implementation !







SPICE Simulation Results

- Horizontal Link Throughput: 710 Mflits/sec
  - Router Throughput : 1100 Mflits/sec
  - Inter-Core wire (2mm) delay : 125 ps
- Serialized (8:1) Vertical Link Throughput: 2080 Mflits/sec
  - Serialization Throughput: 2500 Mflits/sec
  - TSV delay: 20 ps

- Speed ratio : (710*32)/(2080*4) = 2.73 (and not 8 !)

	Self-Controlled Multiplexer 2:1	Self-Controlled Demultiplexor 1:2	Serializer 4:1	Deserializer 1:4	Serializer 8:1	Deserializer 1:8
Transistor count	130	132	390	396	910	924
Latency	80 ps	70 ps	150 ps	130 ps	220 ps	190 ps
Throughput	2.9 Gflits/sec	3.2 Gflits/sec	2.5 Gflits/sec	2.8 Gflits/sec	2.5 Gflits/sec	2.8 Gflits/sec

Conclusion

- The new technology of 3D-Integration opens a new windows to more and more integration of components
- TSVs are the most promising technology of vertical connection with a high bandwidth and a low power consumption
- Due to the yield reduction, 3D-Integrated Circuits are limited on the number of TSVs to be exploited
- The GALS paradigm is demanded as clock distribution in three dimensions is almost impossible
- Asynchronous Networks-on-Chip help to exploit the whole high bandwidth of vertical links (TSVs)
- Serialization of data of vertical links (TSVs) is a trade-off between cost and performance

Merci …