

Réseau Intégré Tridimensionnel Architecture et Implantation

Hamed (Abbas) Sheibanyrad

Frédéric Pétrot



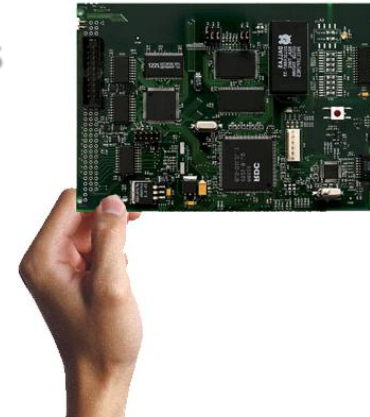
Plan

- L'Intégration Tridimensionnelle
- Distribution de l'Horloge et le Paradigme GALS
- Réseau-sur-Puce Asynchrone
- La Contribution de la Troisième Dimension
- Nouvelles Idées
- Conclusion

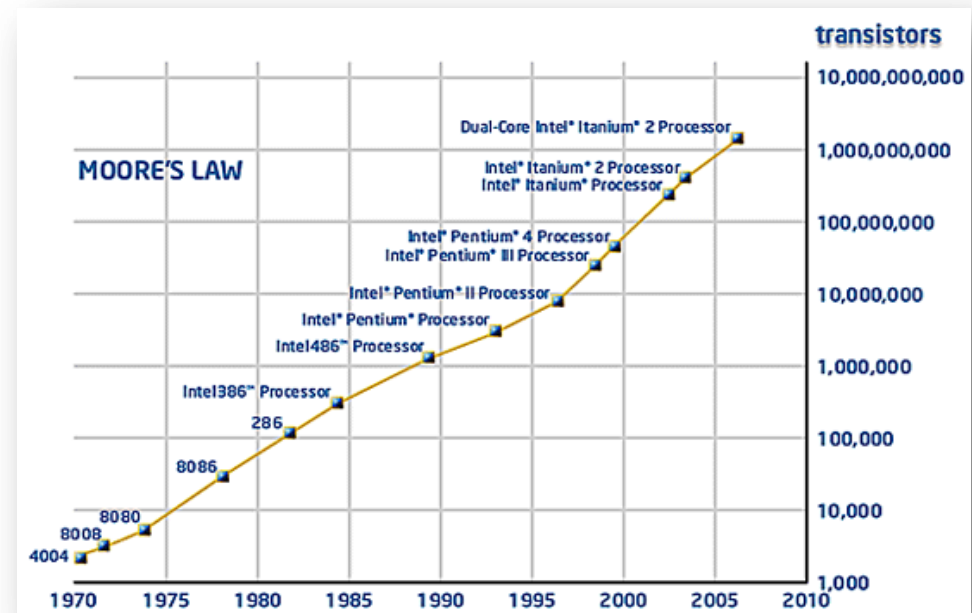
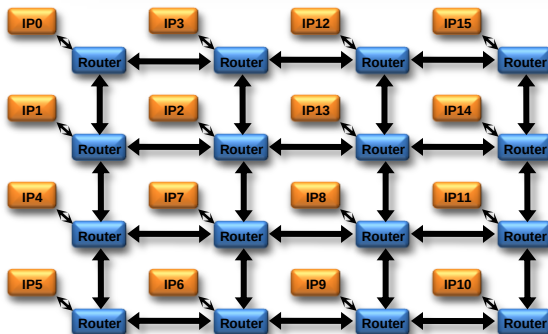
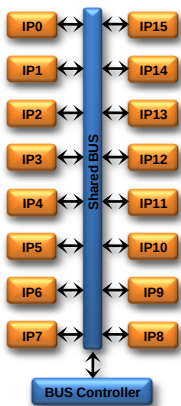
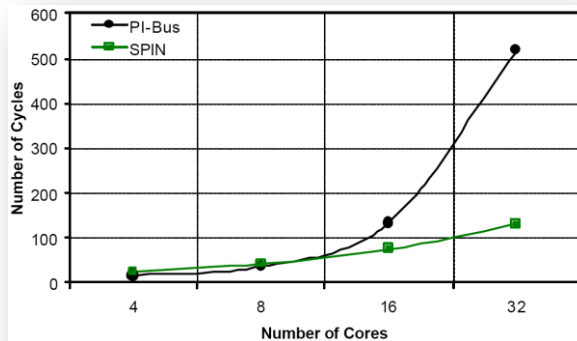
Evolution de la Technologie !

- Evolution de la technologie de fabrication
 - Intégration de système des milliards de transistors dans une seule puce
 - Des centaines et même des milliers de composants
 - Rôle clé de l'infrastructure de communication
 - Passage à l'échelle

Système-sur-Carte

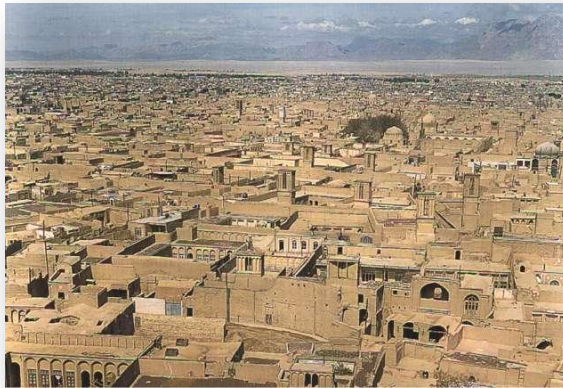


Système-sur-Puce

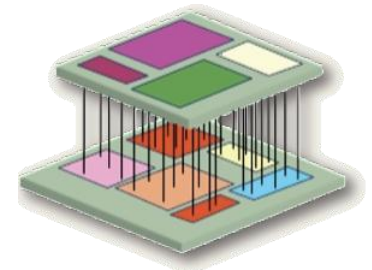
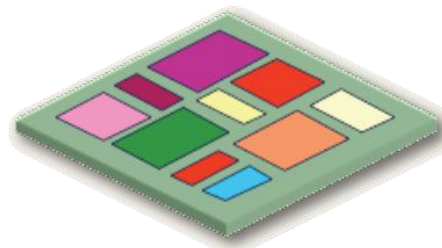


... mais, la terre devient chère !

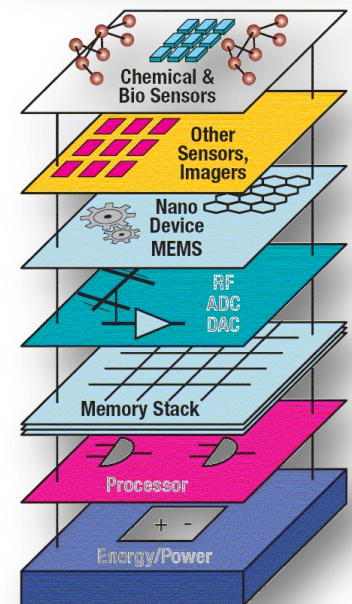
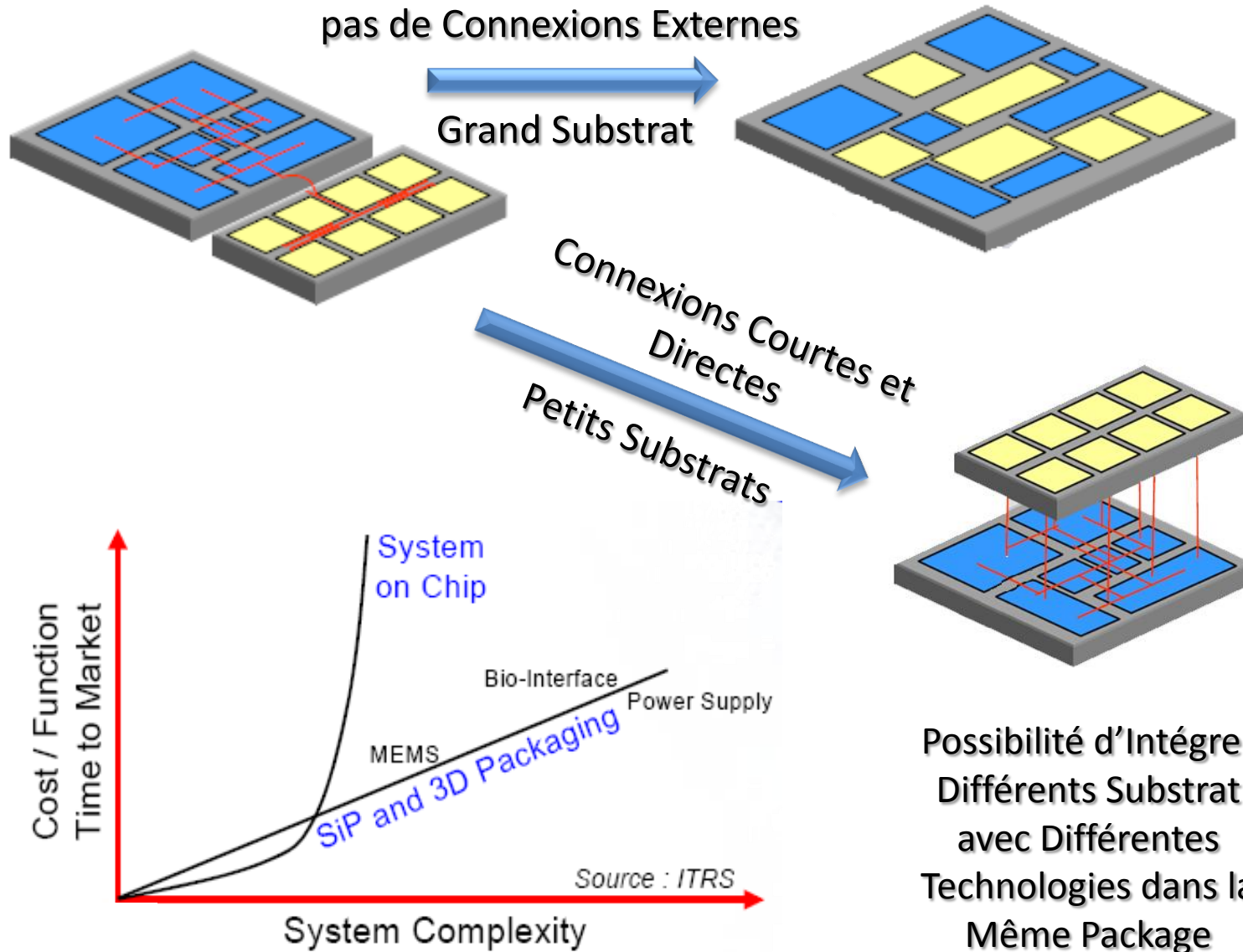
- Comme la terre devient de plus en plus chère, il y a une tendance de construire vers le haut plutôt que l'horizon
 - Augmentation de la densité
 - Réduction de la longueur et du nombre des longs chemins



- Dimensions Limitées du Substrat de Silicium (Die)
 - Contraintes du Rendement (Yield) et la Performance

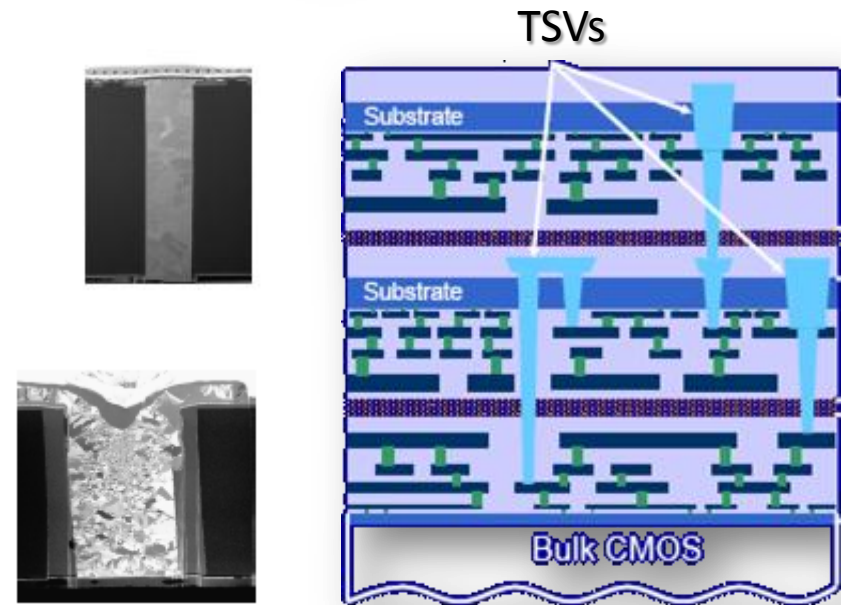
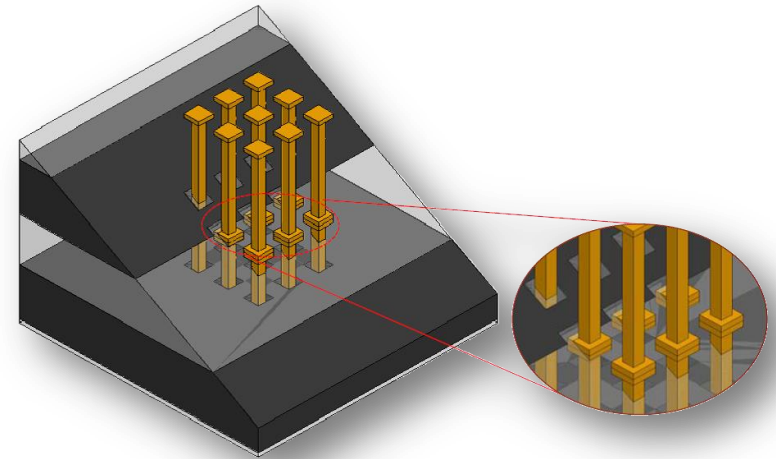


Pourquoi pas l'Intégration 3D des Siliciums ?



Through-Silicon-Via

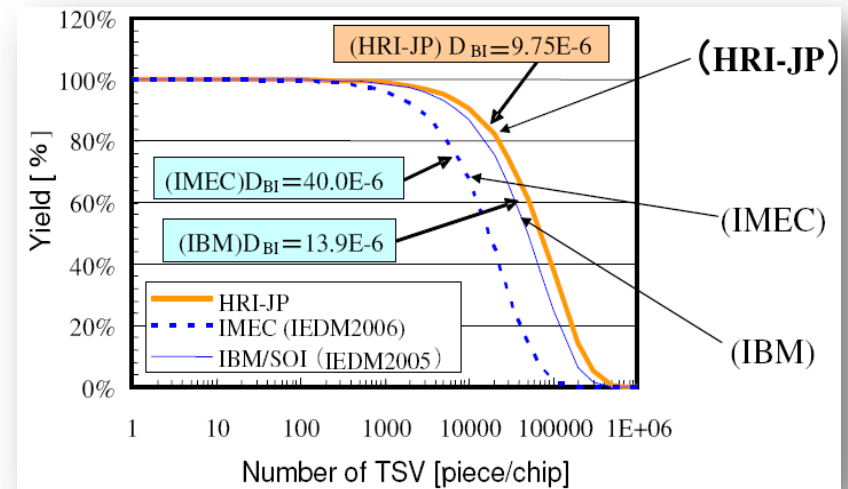
- La Technologie la Plus Prometteuse d'Interconnexion Verticale
 - Résistance et Capacité Faibles
 - Bande Passante Elevée
 - Consommation Faible
- Via-First (densité plus haute des TSVs)
 - Diamètre $\approx 5 \mu\text{m}$ (*IMEC)
 - Pas (**Pitch**) $\approx 10 \mu\text{m}$
 - Profondeur $\approx 20\text{-}50 \mu\text{m}$
- Via-Last (coût plus bas du processus)
 - Diamètre $\approx 35 \mu\text{m}$
 - Pas (**Pitch**) $\approx 60 \mu\text{m}$
 - Profondeur $\approx 40\text{-}150 \mu\text{m}$



... mais, quelle est la Fiabilité des TSVs ?

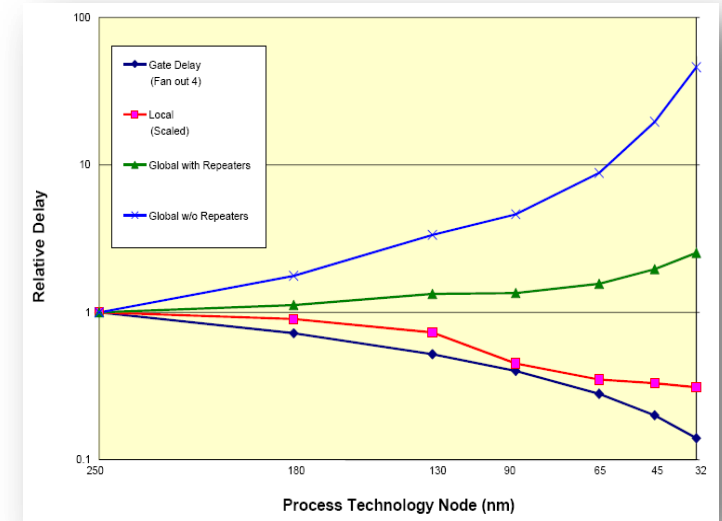
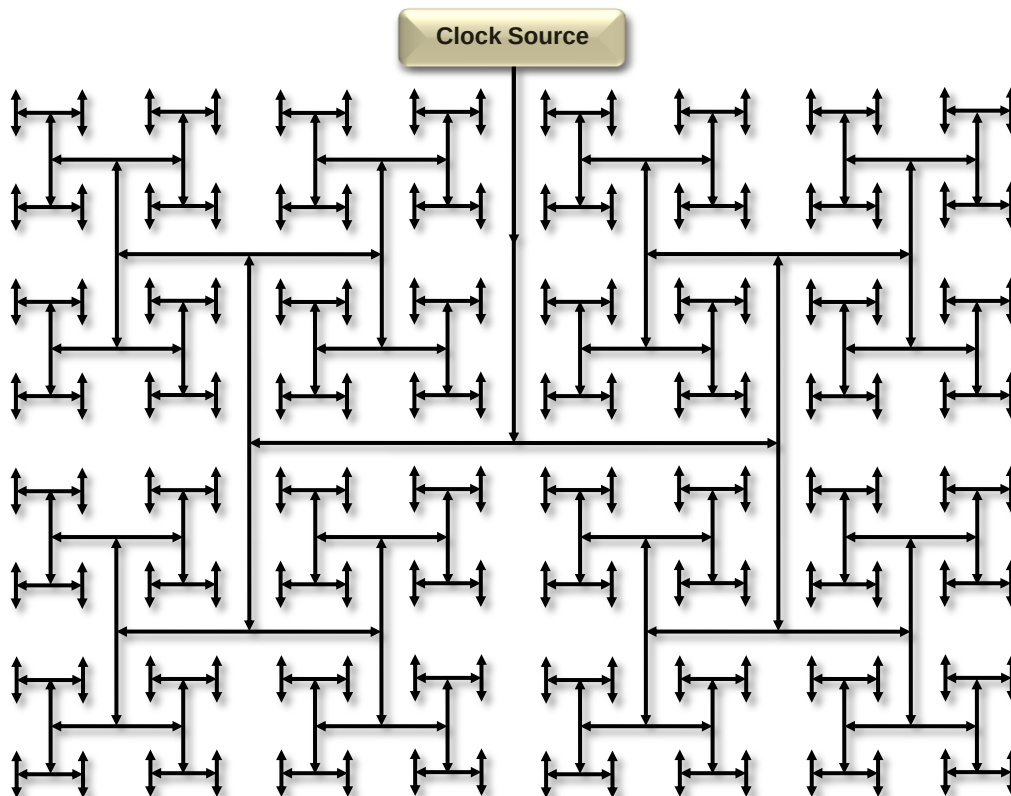
- Grand risque de défaut par nombreuses étapes supplémentaires de fabrication des TSVs (une potentielle réduction du rendement)
 - Désalignement
 - Création des vides pendant la phase de Bonding
 - Dislocation et défauts des grains de cuivre
 - Création des films d'oxyde sur l'interface Cu
 - Détachement partiel ou complet des Pads
 - Défauts liés à la température car la dissipation thermique est de plus en plus difficile
 - ...

Les circuits tridimensionnels
sont limités par le nombre
des TSVs à exploiter



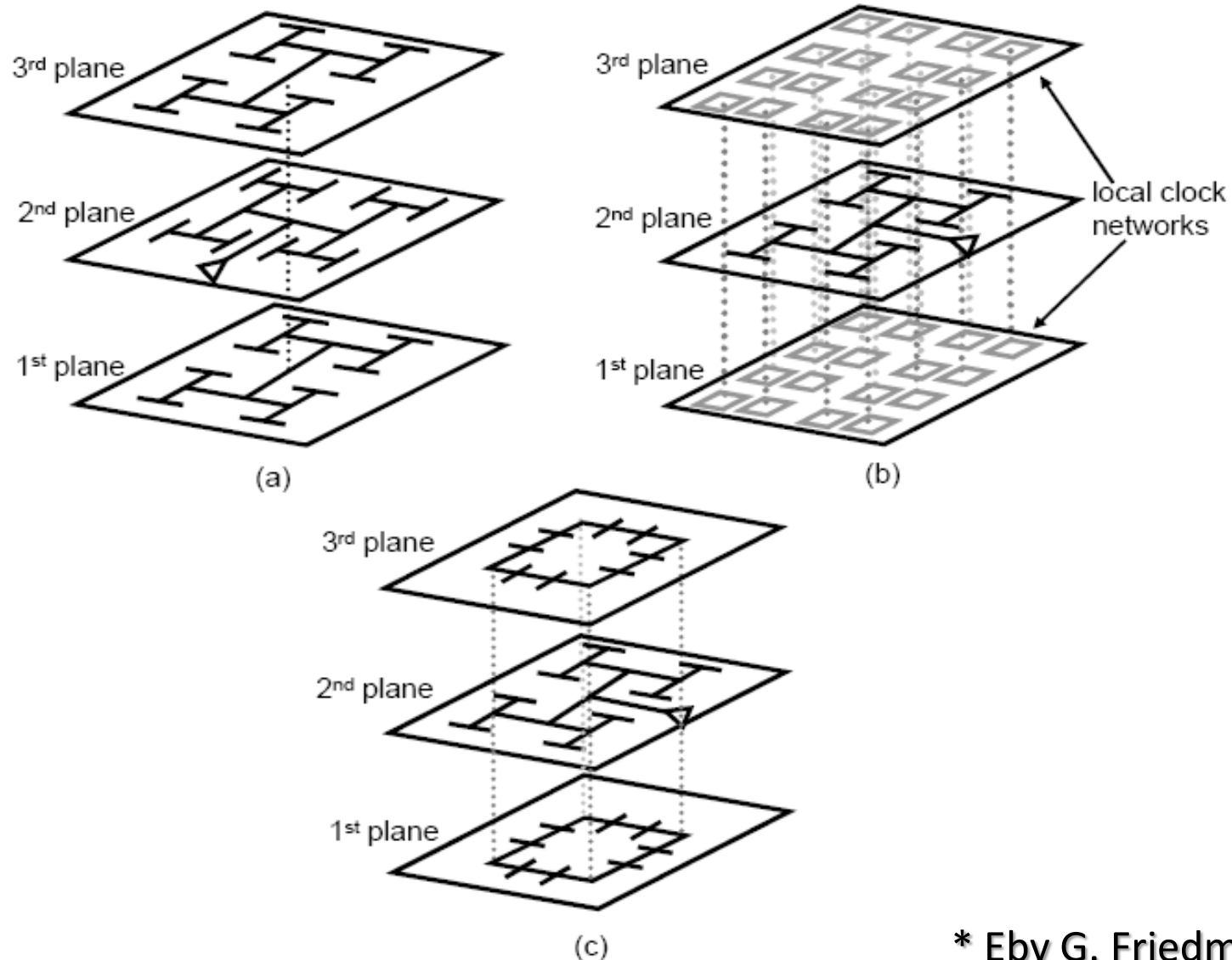
Distribution de l'Horloge

- Technologies Submicroniques
 - Aggravation des problèmes physiques
 - Effet des fils sur le délai et la consommation prédominant



- Cauchemar de la synchronisation globale
 - Distribution globale d'un seul signal d'horloge sur une puce impossible
 - Variation de processus de fabrication
 - Variation de température

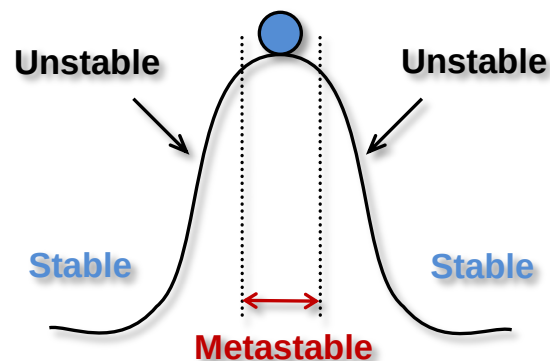
Distribution de l'Horloge en 3 Dimensions



* Eby G. Friedman

GALS toujours Exigé !

- Le paradigme de GALS (Globalement Asynchrone Localement Synchrone) est une solution attractive
 - Plusieurs domaines cadencés indépendamment les uns des autres
- Les Réseaux-sur-Puce sont les approches les plus structurées
 - Le Réseau est la partie globalement asynchrone du système
 - Les sous-systèmes sont les parties localement synchrones du système

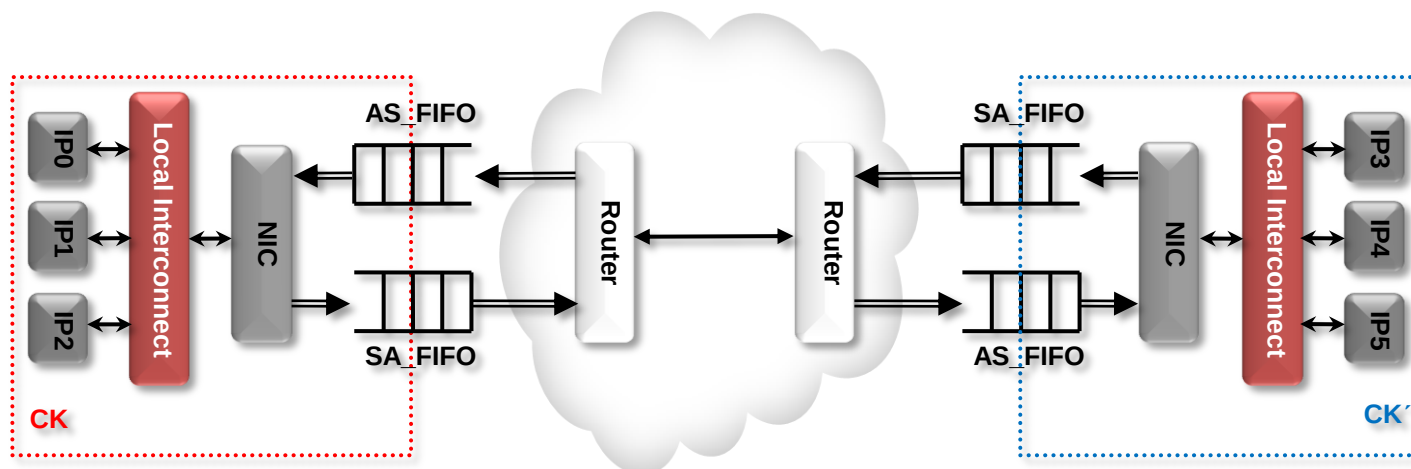


... **mais**, comment deux domaines cadencés séparément, peuvent-ils communiquer de manière fiable ?

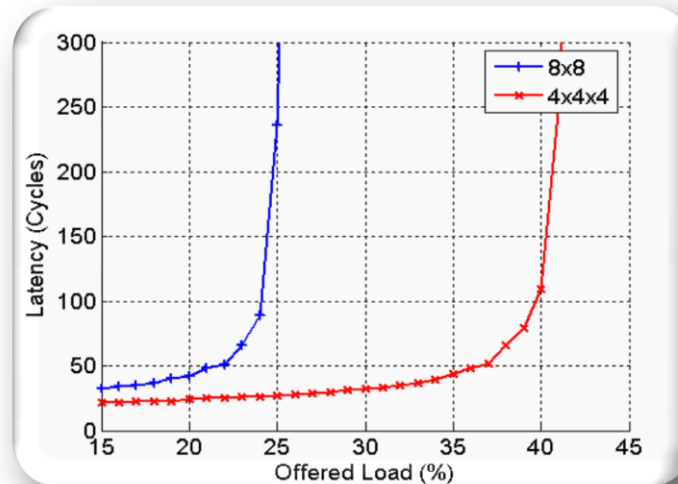
- La Métastabilité, un état inévitable d'un système bistable, est le problème majeur des architectures GALS

ASPIN: un Réseau Entièrement Asynchrone

- Le besoin de synchronisation réduit aux interfaces du réseau
 - des FIFOs Spécifiques: Async-to-Sync et Sync-to-Async
 - Une latence bout-à-bout beaucoup plus faible que la version multi-synchrone
- Le plus rapide possible et indépendant du reste du circuit
 - Un seuil de saturation amélioré par rapport à la version multi-synchrone
- La Consommation Dynamique Quasi-Zéro à l'état de repos
- Scalabilité et Réutilisabilité de manière Prêt-à-Tourner (Plug & Play) et indépendant de la taille des sous-systèmes



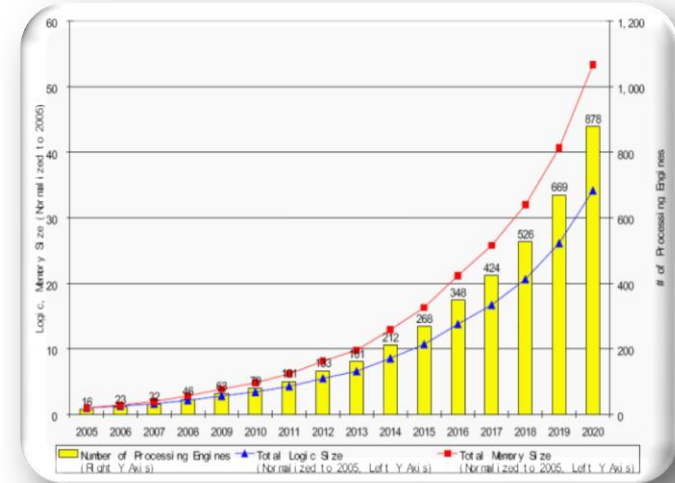
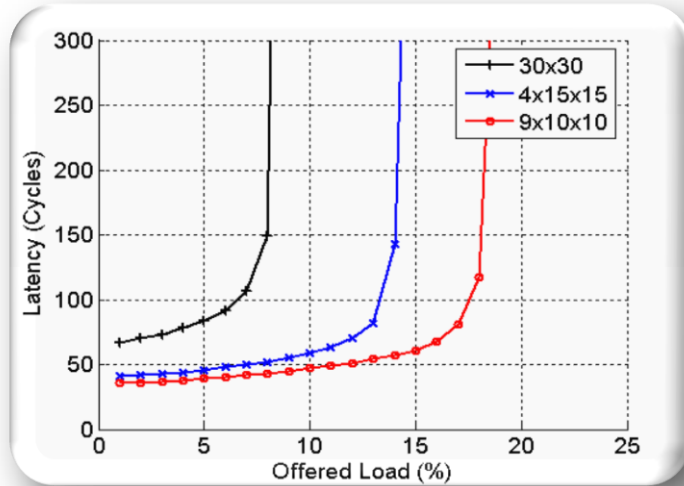
Pourquoi la Troisième Dimension ?



	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

Pourquoi la Troisième Dimension ?

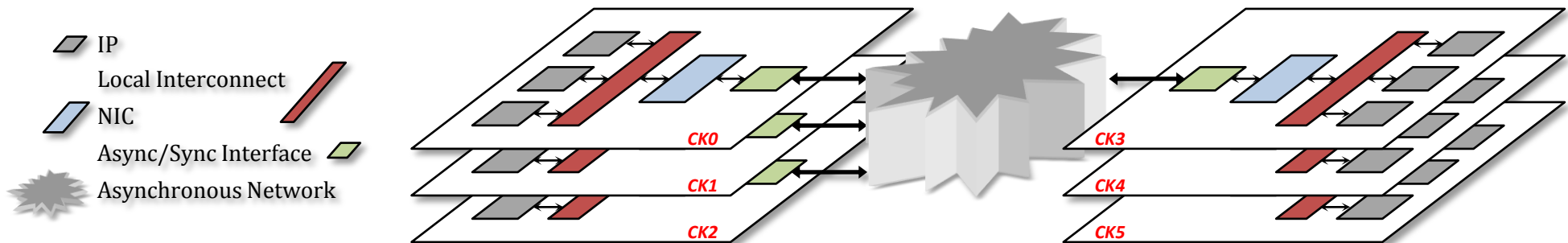
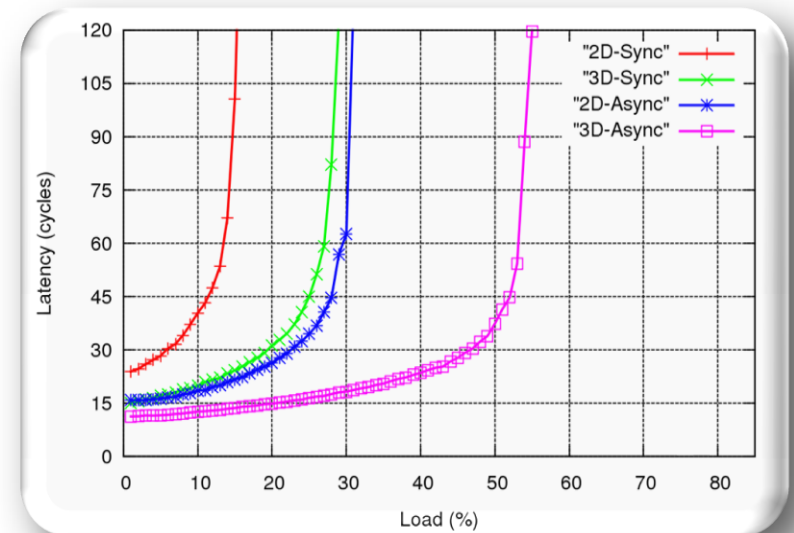


	Number of Nodes	Switch Degree	Network Diameter	Number of Channels	Number of Vertical Channels	Number of Bisection Channels	Load of the Busiest Channels ⁽¹⁾
30x30	900	5	60	5280	0	60	$C \times \frac{1}{4} \times 30$
4x15x15	900	7	34	6510	1350	120	$C \times \frac{1}{4} \times 15$
9x10x10	900	7	29	6640	1600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

Réseau Asynchrone 3D

- Insensible à la Variation de Délai par suite de la Variation de Température et la Variation de processus
- Exploiter la totalité de la Bande Passante Elevée des TSVs
- Rapport de vitesse de 2 comme le pire des cas
 - Exploitant les transistors 90nm GPLVT de STMicroelectronics, 400MHz comme la fréquence la plus élevée des SoCs usuels
 - Exploitant la même technologie, 1100 Mflits/s comme la bande passante d'un NoC asynchrone

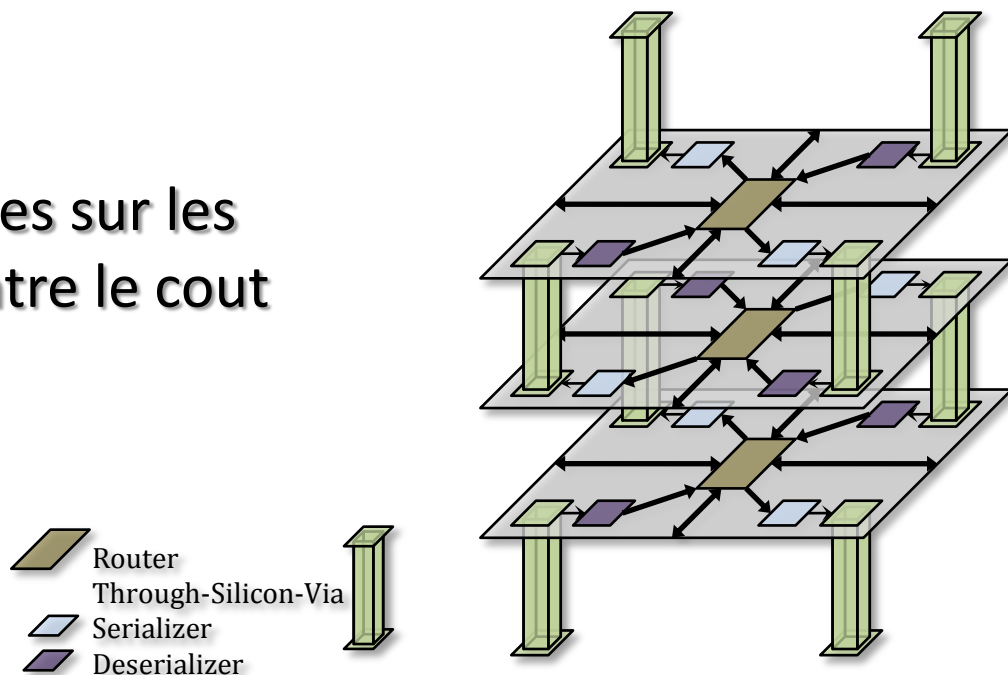


Liens Verticaux Séries !

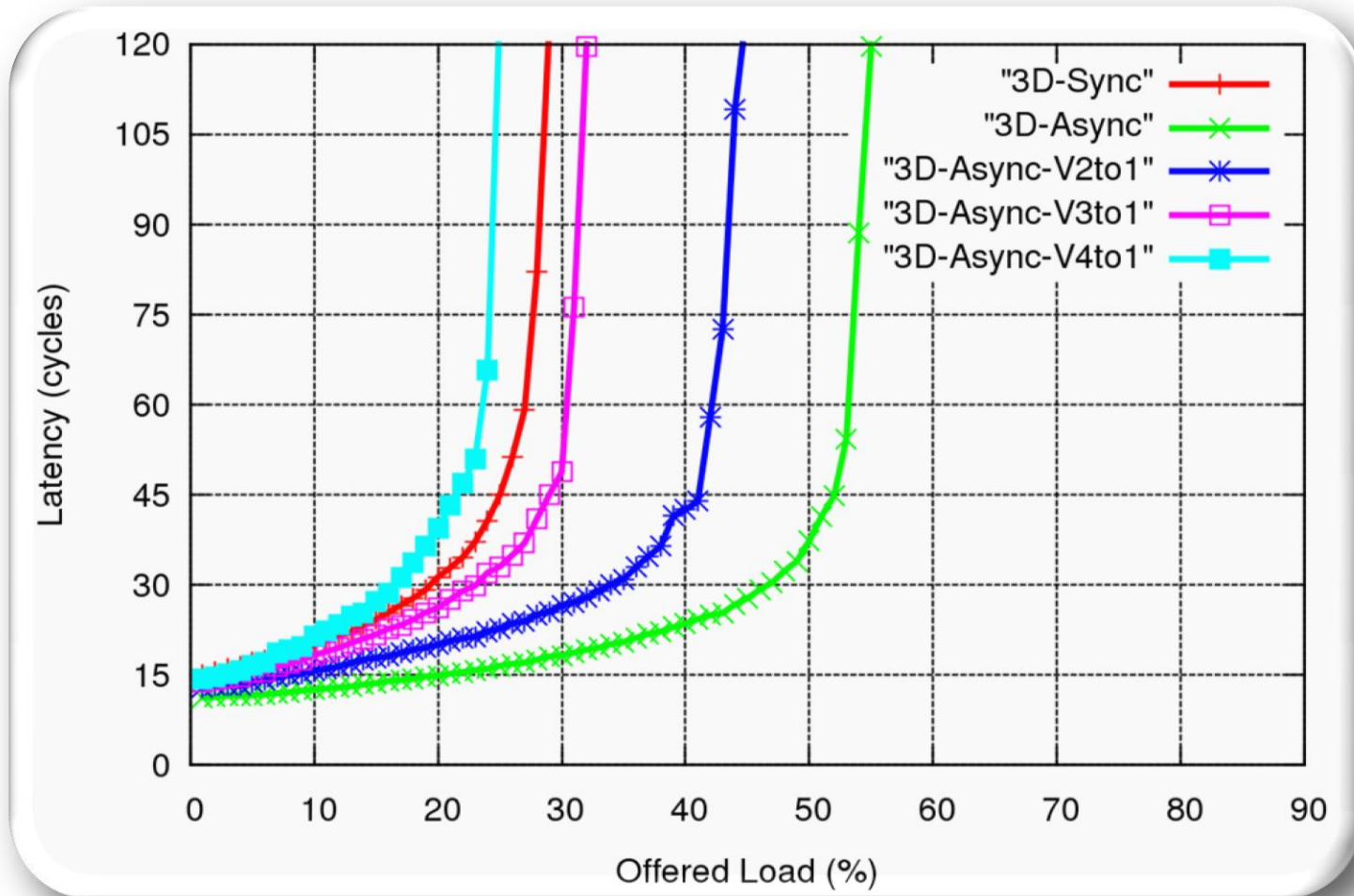
- Rappel

- L'utilisation des TSVs garantit un échange de données vertical plus rapide et moins consommant que les liaisons horizontales de tailles modérées
- mais, le pas (Pitch) des TSVs est grand et impose un surcout inacceptable en surface, et, les nombreuses étapes supplémentaires de la fabrication des TSVs ajoutent une potentielle réduction du rendement (Yield)

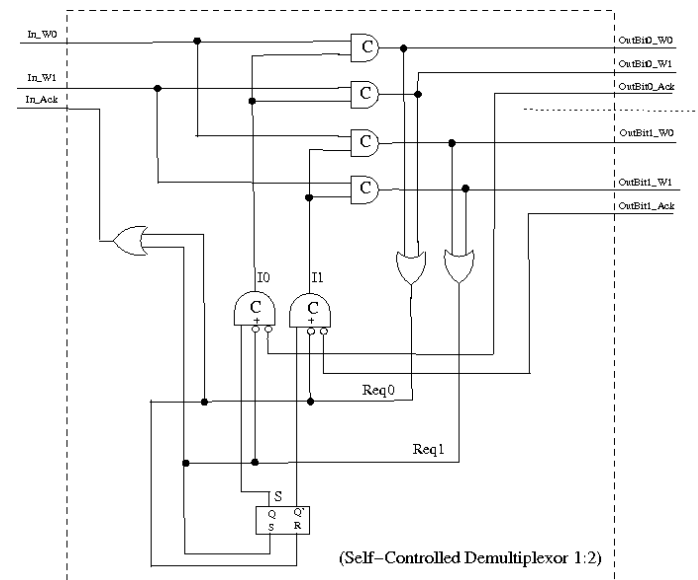
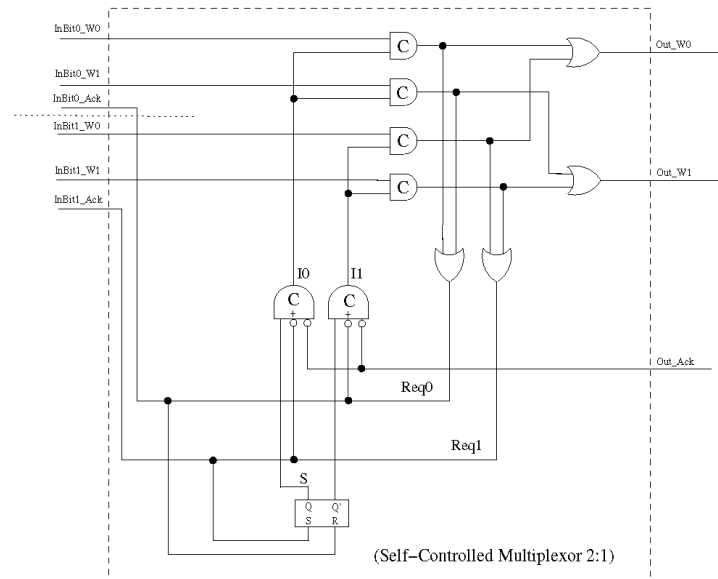
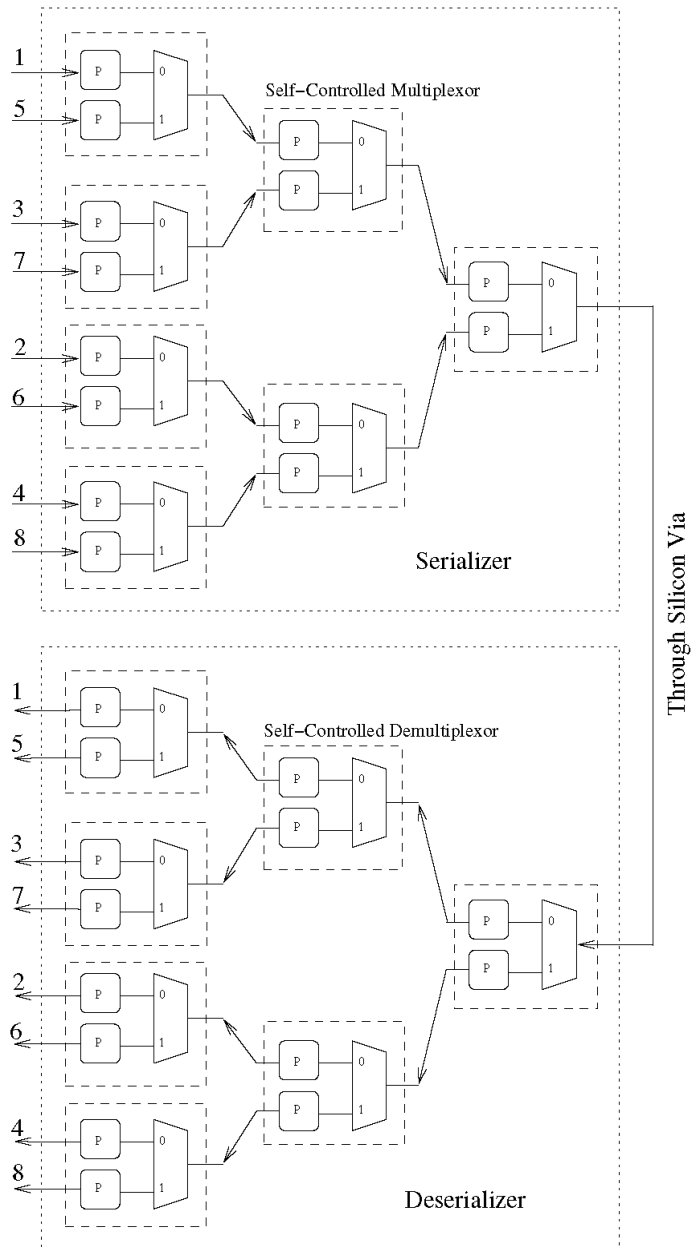
- La sérialisation des données sur les TSVs est un compromis entre le cout et la performance



NoC Asynchrone 3D Verticalement Série



Implémentation du Circuit !



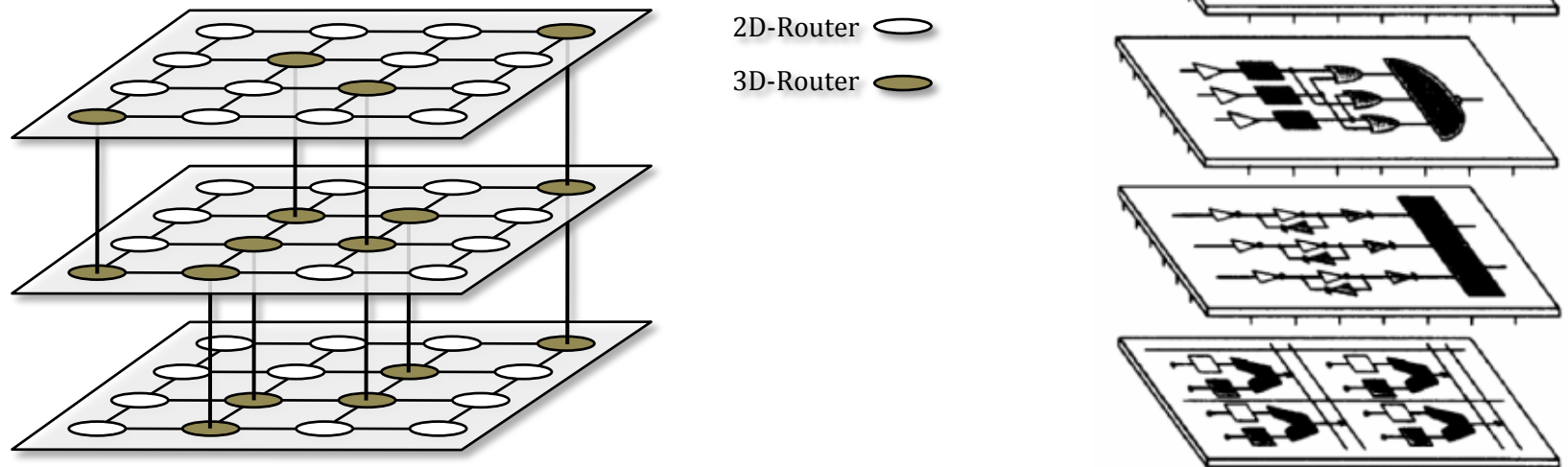
Résultats de Simulation SPICE

- Bande Passante des Liens Horizontaux : 710 Mflits/sec
 - Bande Passante de Routeur : 1100 Mflits/sec
 - Délai de Fil d'Inter-Cluster (2mm) : 125 ps
- Bande Passante des Liens Verticaux Séries (8:1) : 2080 Mflits/sec
 - Bande Passante de Sérialisation : 2500 Mflits/sec
 - Délai de TSV : 20 ps
- Rapport de Vitesse : $(710*32)/(2080*4) = 2.73$ (et pas 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

Connexion Verticale Partielle !

- Nombre de connexions verticales (TSVs) limité
- Réseau incluant différent substrats et avec différent technologies
 - Hétérogénéité
 - Non Régularité
- Une topologie verticalement partiellement connectée est une solution efficace
 - Mécanisme de routage dans une telle topologie irrégulière est le problème majeur



Conclusion

- La nouvelle technologie de l'intégration 3D ouvre une nouvelle fenêtre sur l'intégration de plus en plus de composants
- TSVs sont la technologie la plus prometteuse de connexion verticale avec une bande passante très élevée et une consommation faible
- Les circuits tridimensionnel sont limités sur le nombre des TSVs à exploiter à cause du rendement
- Le paradigme GALS est exigé car la distribution de l'horloge en trois dimensions est quasiment impossible
- Réseau-sur-Puce Asynchrone aide à exploiter la totalité de la bande passante des liens verticaux (TSVs)

Conclusion

- La contribution de la troisième dimension dans l'architecture d'un réseau aide à améliorer la performance du système
- La sérialisation de l'échange de données sur les liens verticaux (TSVs) est un compromis entre la cout et la performance
- Une topologie verticalement partiellement connectée est une solution efficace pour diminuer le nombre des TSVs et s'adapter à la hétérogénéité et la non-régularité des systèmes intégrés tridimensionnels

Une Bonne Nouvelle !

- Edition d'un Ouvrage
 - Abbas Sheibanyrad (TIMA), Frédéric Pétrot (TIMA), Axel Jantcsh (KTH), **3D-Architectures and Networks-on-Chip**, Morgan Kaufmann (éventuellement !)

Merci ...