

# Chiplet integration on active silicon interposer

Applied Power Electronics Conference – Invited talk | 2021, June 9-12<sup>th</sup>

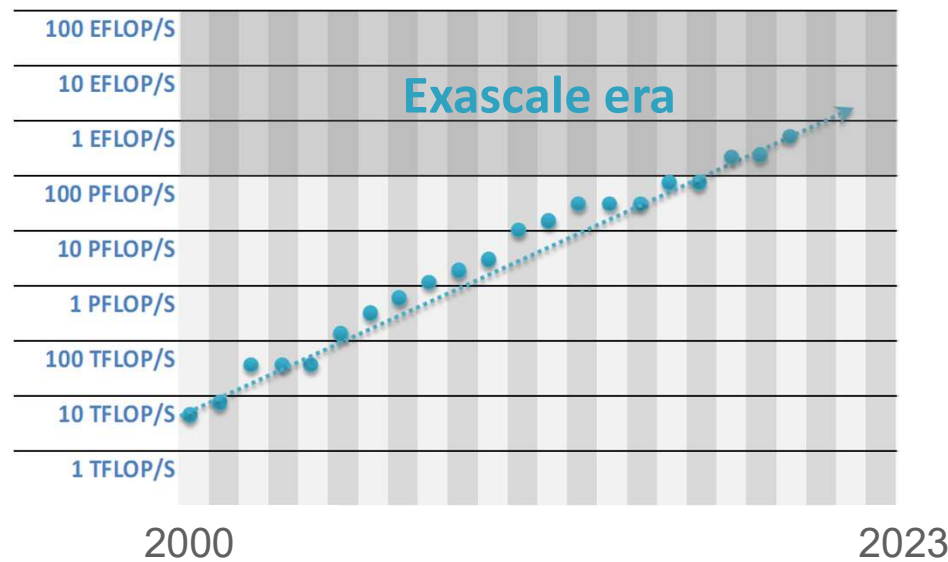
Perceval Coudrain, S. Chéramy, E. Ollier, J. Charbonnier, E. Bourjot, D. Dutoit, P. Vivet, F. Clermidy  
CEA – LETI, CEA – LIST, Grenoble, France

# Agenda

- **Exascale computing trends**  
Needs for heterogeneous 3D integration
- **INTACT Demonstration**  
Many-core integration on active silicon interposer
- **Towards ultra-fine pitch interconnects**  
Hybrid bonding approaches

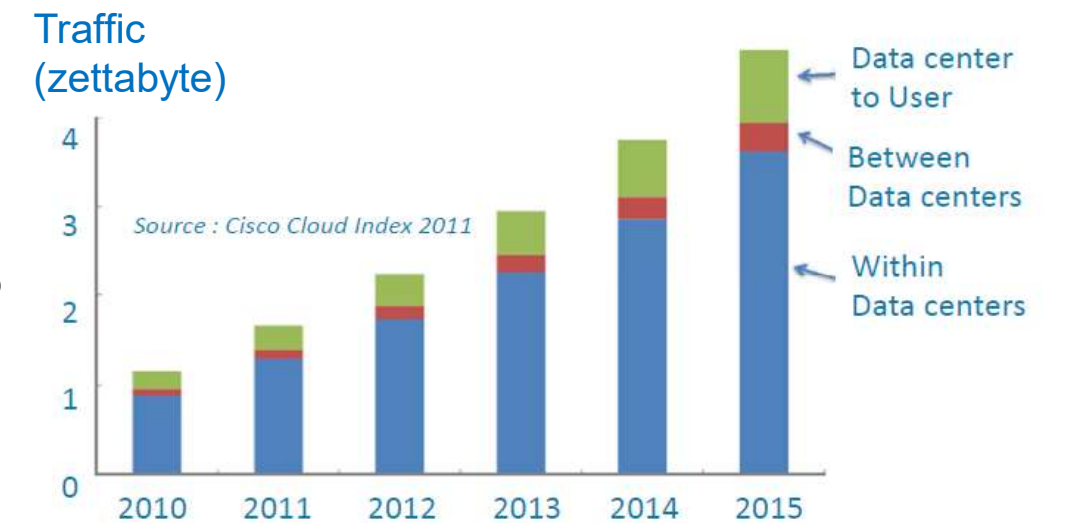
# Challenges in computing & big data applications

## Performance



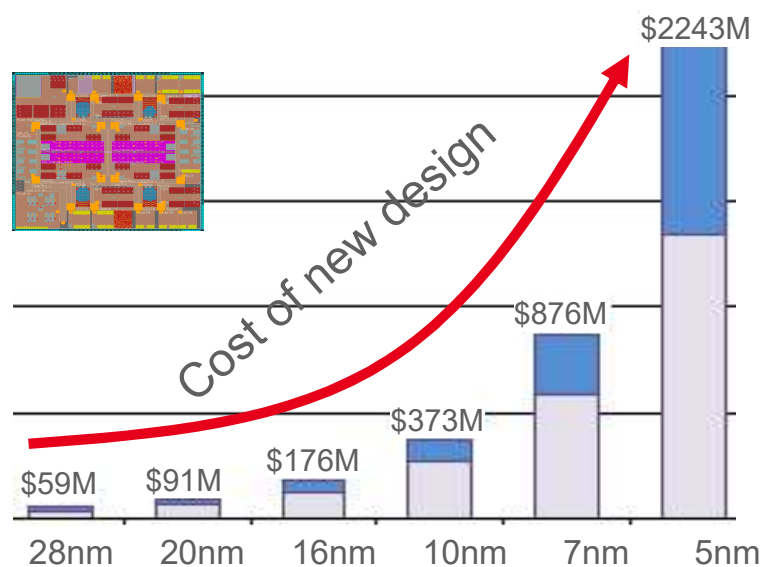
$\times 10$   
/4 years

## Data Traffic



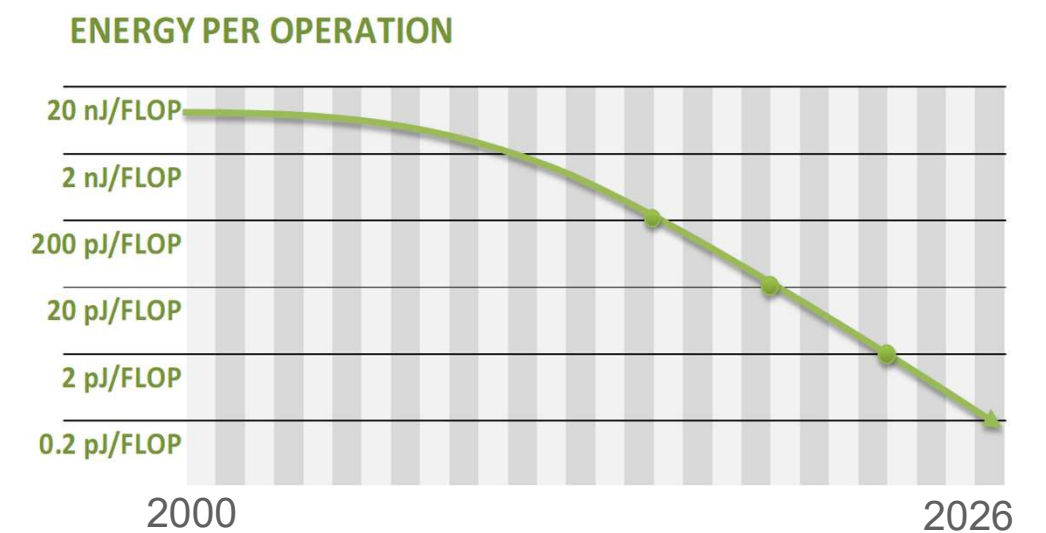
$\times 2$   
/2 years

## Cost



$\times 2$   
/CMOS  
node

## Energy per Operation

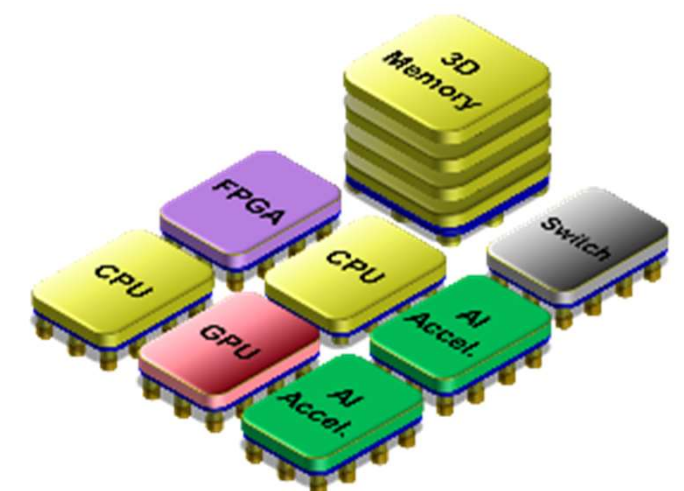
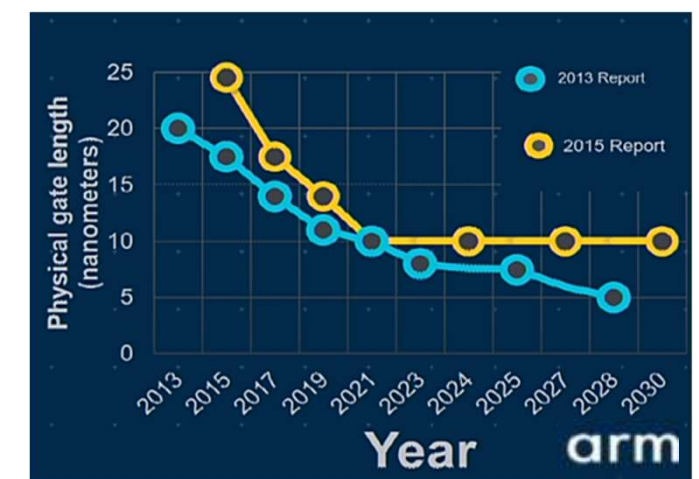
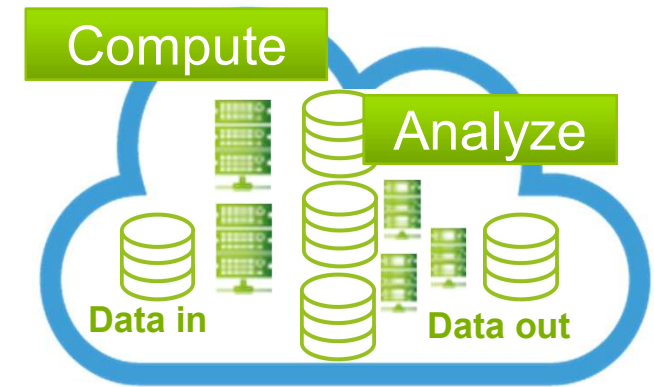


/ 4  
/2 years

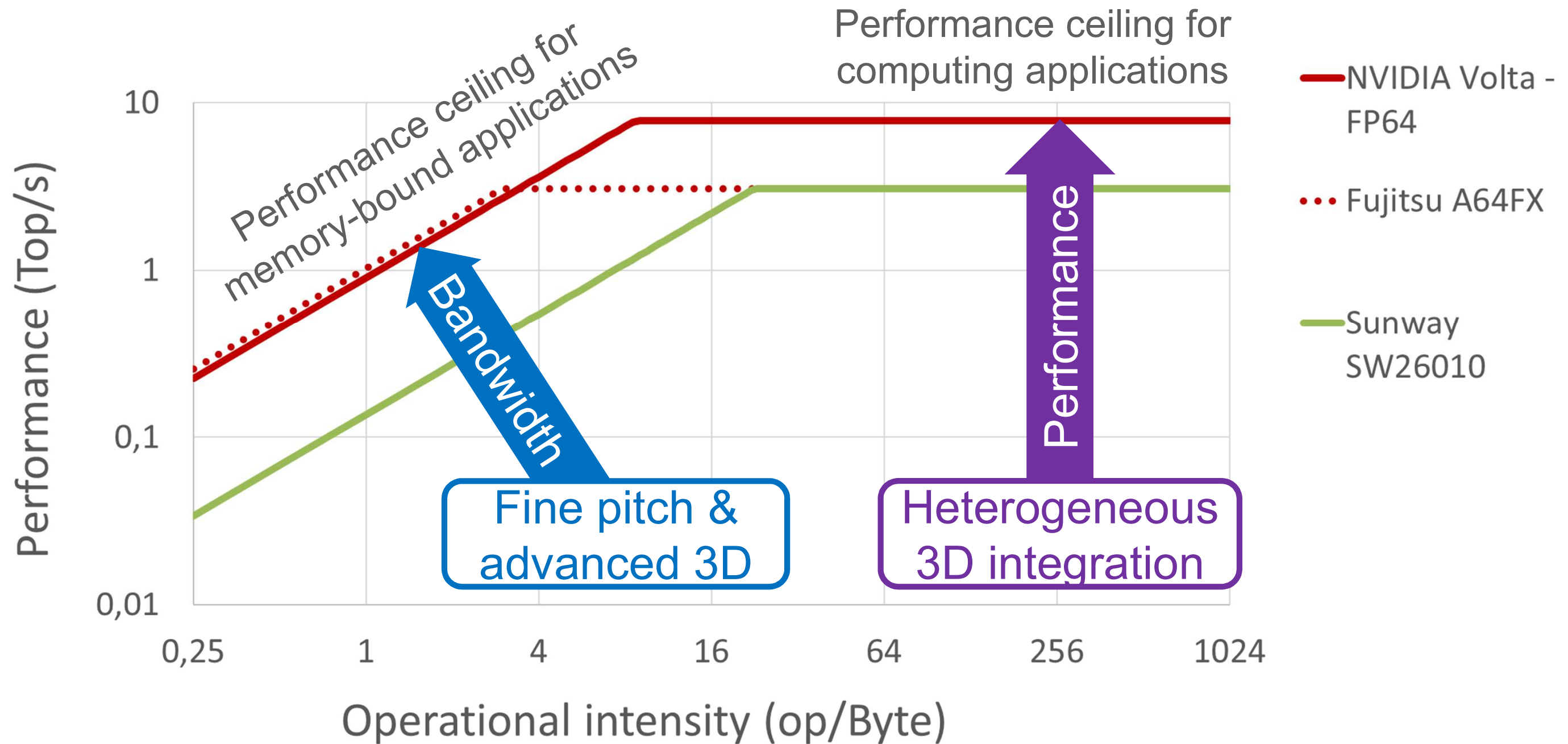


# High performance computing (HPC) & Big data

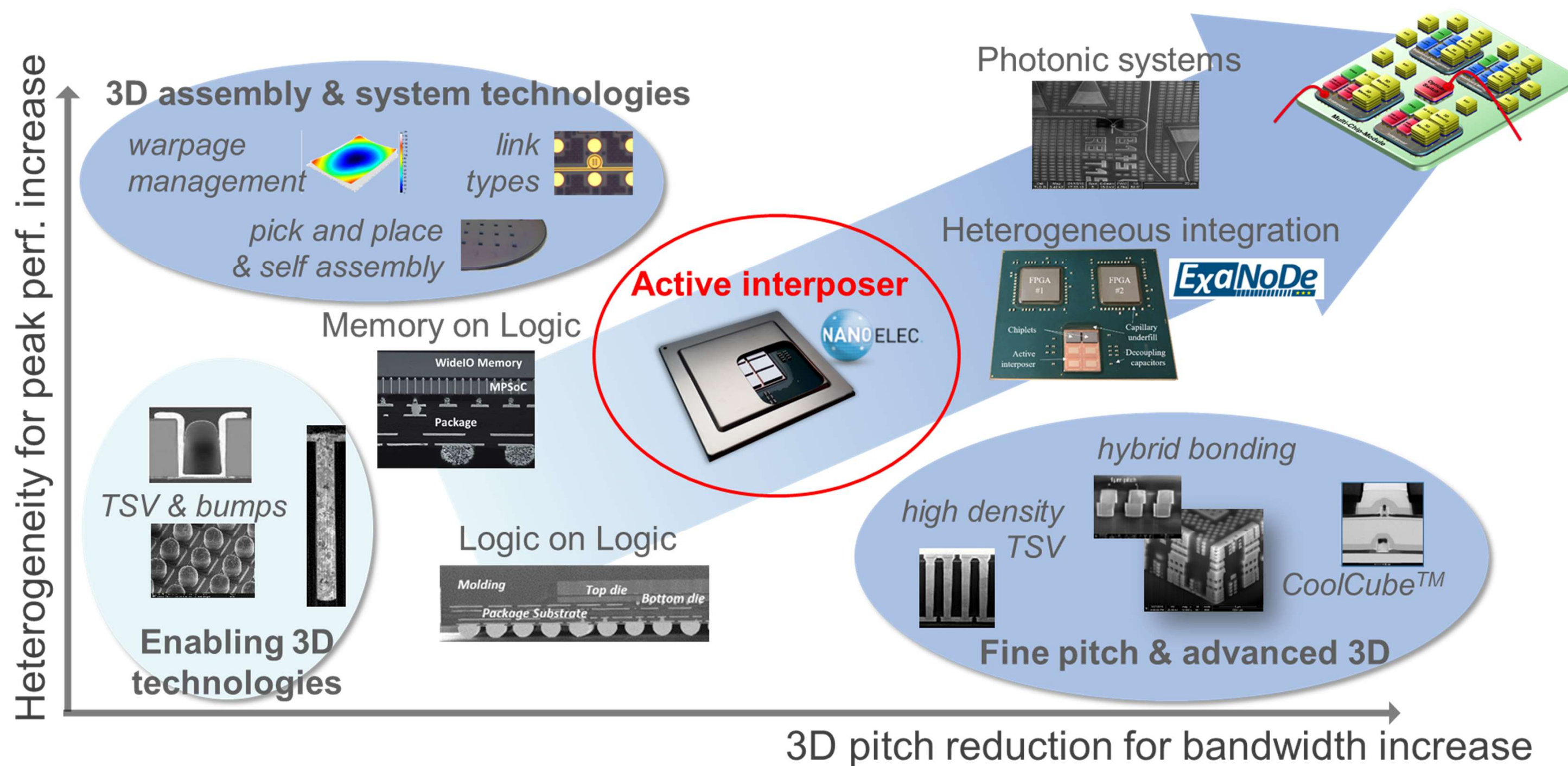
- **More cores, more accelerators, more memory**
  - Highly optimized **generic** & **specialized** functions
  - Go-to-market solution for **sustainable system differentiation**
- **Single-die SoC approach not viable anymore**
  - **Digital** → advanced CMOS cost & yield
  - **Analog** → does not shrink anymore
- **System designers must offer**
  - Modular and cost effective solutions
  - **Energy efficiency** in system infrastructure
  - More on-chip memory bandwidth per core



# Roofline model for performance comparison



# 3D integration roadmap for HPC





# Why considering chiplets on active interposer ?

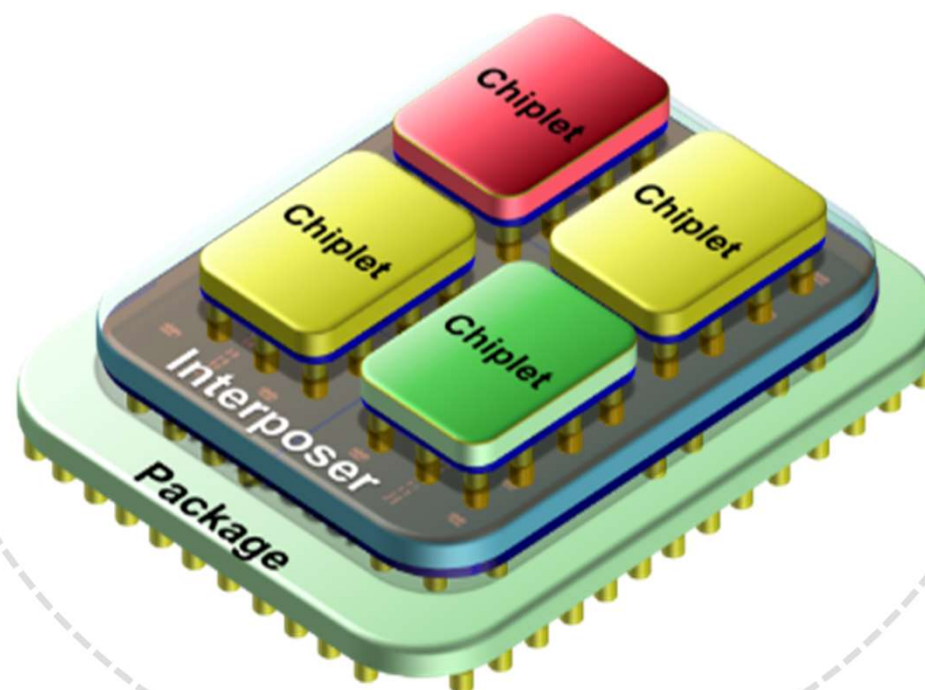
## Improved cost / performance tradeoff

- Small to medium size chips
- Advanced technology node only when needed
- Function-partitioning

## Specialization

- Heterogeneous integration
- Flexible communication between chiplets
- Asynchronous network on chip (NoC) for latency reduction

## Scalable computing component



## Improved integration

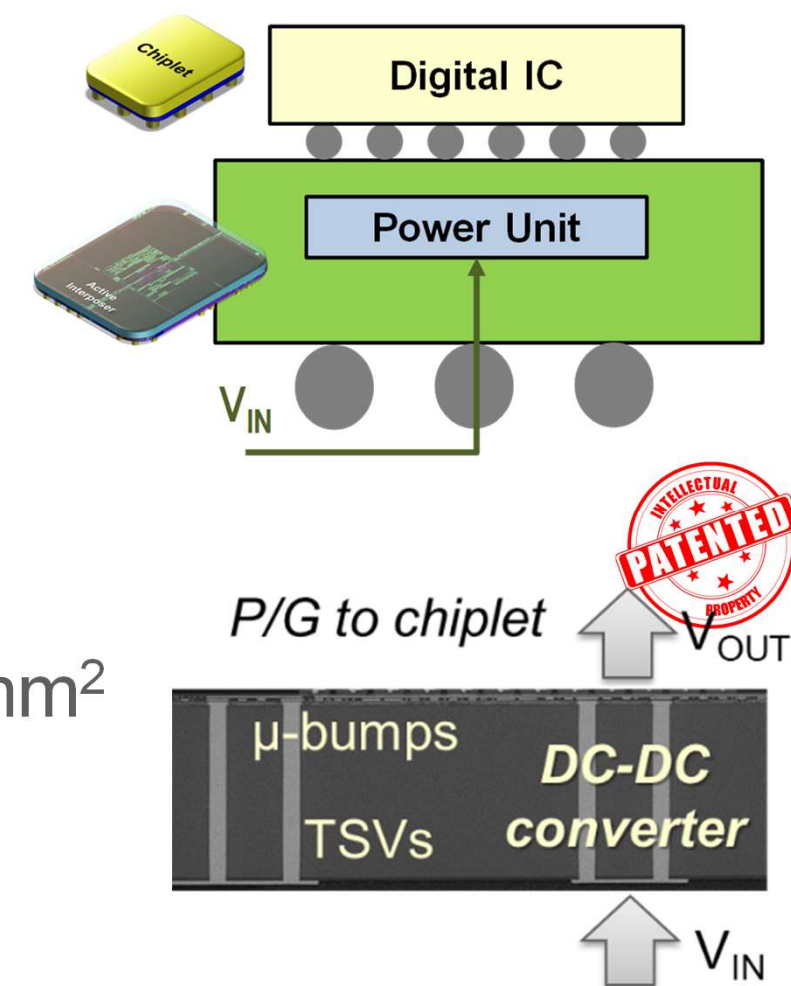
- Scale-out concept for many-core architectures
- Scalable cache-coherent memory hierarchy
- Adaptation between heterogeneous technologies

## Improved energy & thermal management

- Energy efficient voltage regulator close to computing chiplet
- Embedded Sensors

# Focus on switched cap voltage regulators (SCVR)

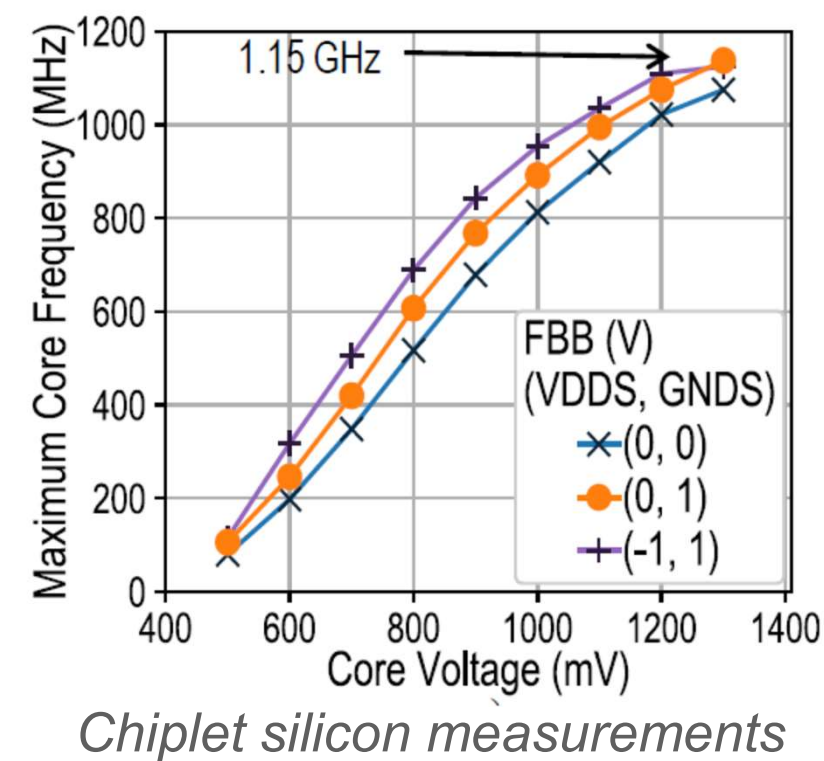
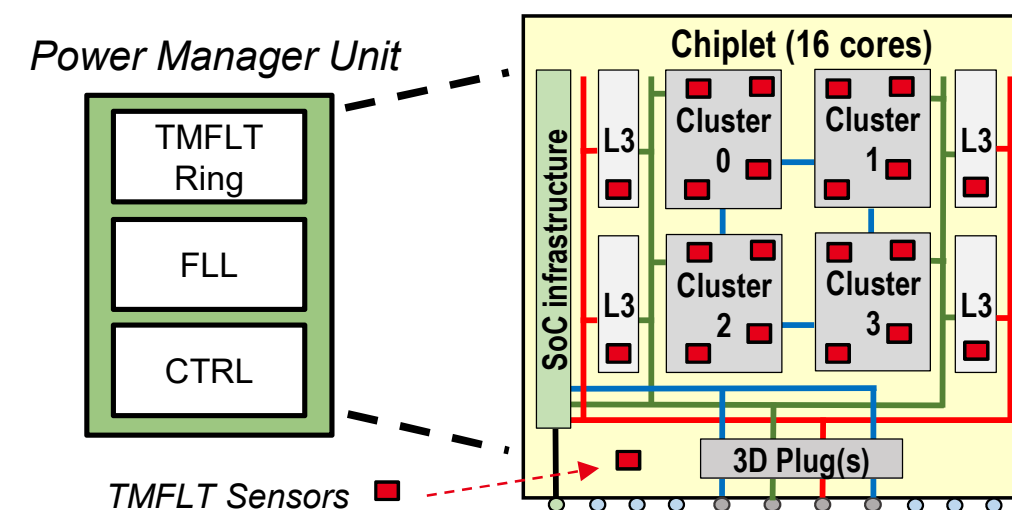
- **Distributed power supply units**
    - Local dynamic voltage & frequency scaling (DVFS)
    - Fast transitions & reduced IR-drop effects
    - High input voltage (up to 2.5V) reduces # power grid IOs
  - **Full integration in interposer**
    - No external passives, on-chip MOS+MOM+MIM caps → 8.9 nF/mm<sup>2</sup>
    - 50% of chiplet area, power grid delivery as μ-bump flip-chip array
  - **Power Management with  $f_{\max}/V_{\min}$  tracking**
    - Circuit estimates  $V_{\min}$  and  $f_{\max}$  & adapts V or F to track optimal energy point
    - Time Fault Sensor (*Canary FF like*) estimates  $f_{\max}/V_{\min}$  of circuit during calibration phase
    - Time Fault Ring (*configurable replica path*) tracks optimal  $f_{\max}$  or  $V_{\min}$  along circuit life-time
- 
- The diagram illustrates the system architecture and power management components. At the top, a yellow box labeled "Digital IC" is connected via several small circles to a green box labeled "Power Unit". To the left of the Power Unit are two smaller chips: a yellow one labeled "Chiplet" and a blue one labeled "Active Interposer". Below the Power Unit, three gray circles represent input pins, with the first one labeled  $V_{IN}$ . An arrow points from the  $V_{IN}$  pin up to the Power Unit. Below the Power Unit is a dark gray rectangular block representing the interposer. Inside this block, there are vertical columns labeled " $\mu$ -bumps" and "TSVs". To the right of these columns, there is a red circle with a white "P" and a label "DC-D converter". An arrow labeled "P/G to chiplet" points from the interposer block up towards the Chiplet and Active Interposer. Another arrow points upwards from below the interposer block.





# Power management benefits of $f_{\max}/V_{\min}$ tracking

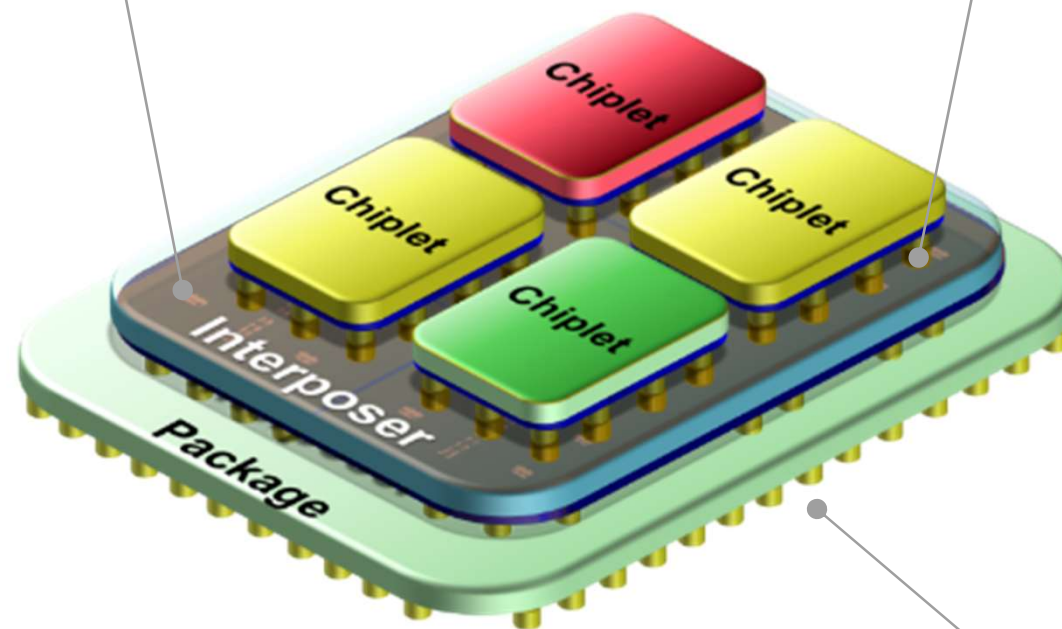
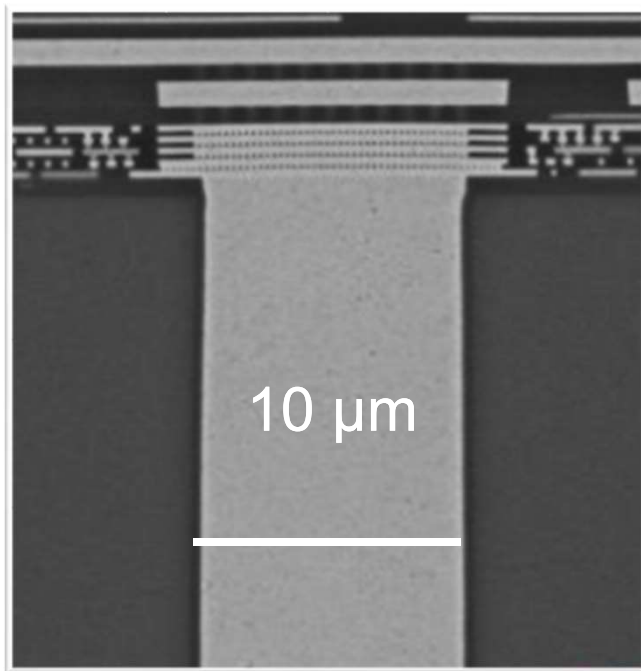
- **PMU integration within INTACT chiplets**
  - 1024 Time Fault Sensors per chiplet, 93% capturing useful information → 6X w.r.t. other solutions
  - 4 Time Fault Rings with 8mV tracking sensibility
  - Controller adapts  $f_{\text{clock}}$  to track PVT variations & delay
- **Tradeoff  $f_{\max}$  versus  $V_{\max}$** 
  - $V_{\min}/F_{\max}$  estimated on wide voltage range [0.4 V – 1.3 V]
  - Estimation error ~2% over full voltage range
  - Power gain from 19% to 36% w.r.t. signoff CAD results
- **Achieved with and without FDSOI Body Biasing**
  - Body biasing range [0;0], [0;+1V], [-1V;+1V]



# Active silicon interposer technology

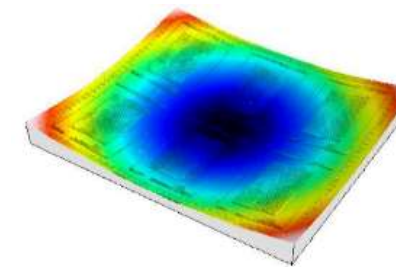
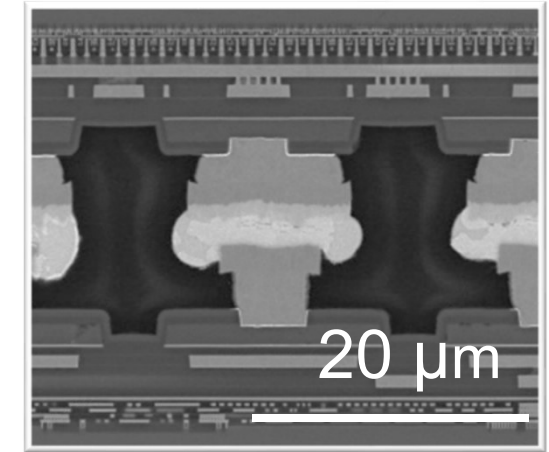
## Through silicon vias (TSV) middle

- Diameter 10 $\mu$ m
- Depth 100 $\mu$ m (AR 10:1)
- Pitch 40 $\mu$ m



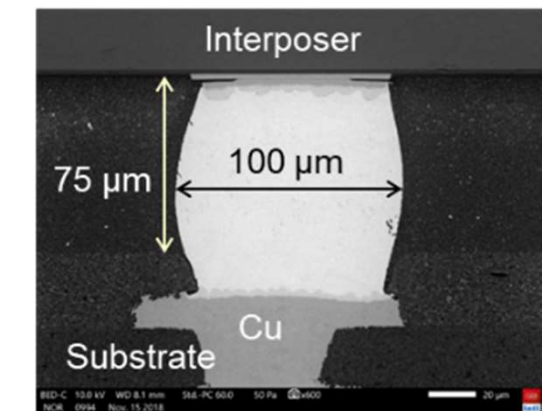
## $\mu$ -bumps

- Pitch 20 $\mu$ m
- Diameter 10 $\mu$ m
- 150 000 interconnects on interposer (25k/chiplet)



## Packaging

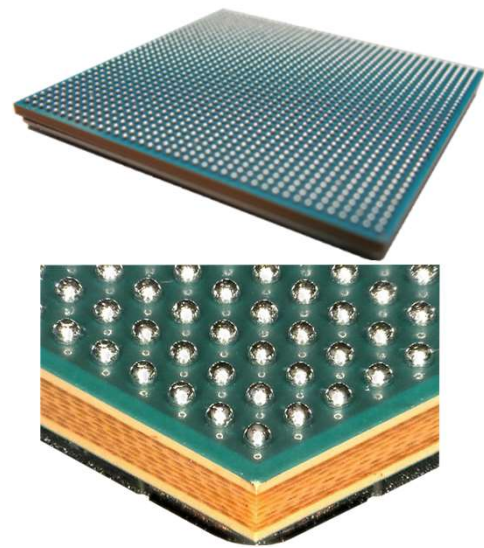
- Warp management
- Alignment accuracy



# INTACT active interposer demonstration



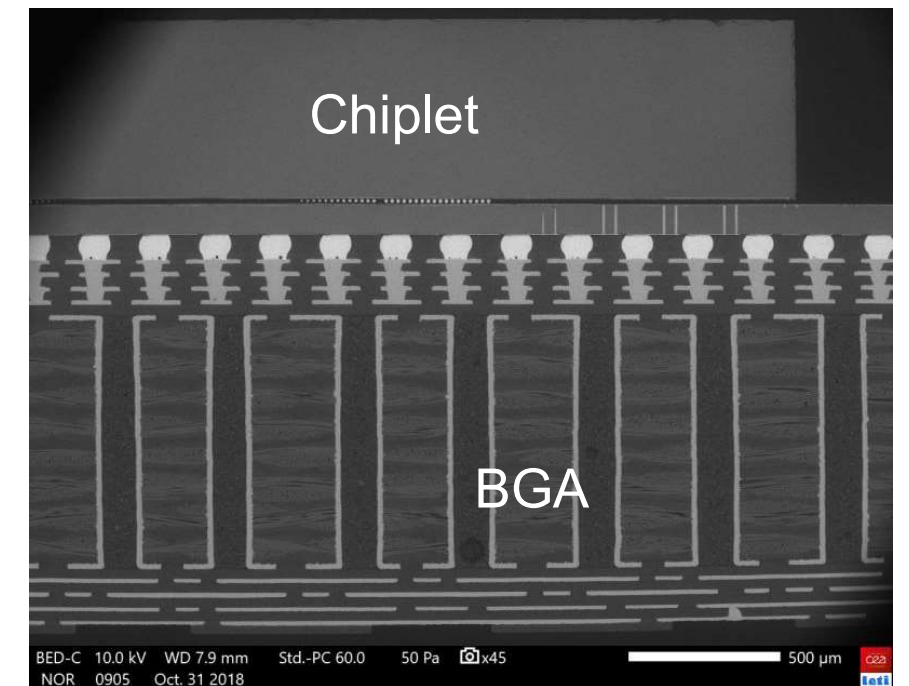
*INTACT prototype*



*INTACT  
BGA after balling*

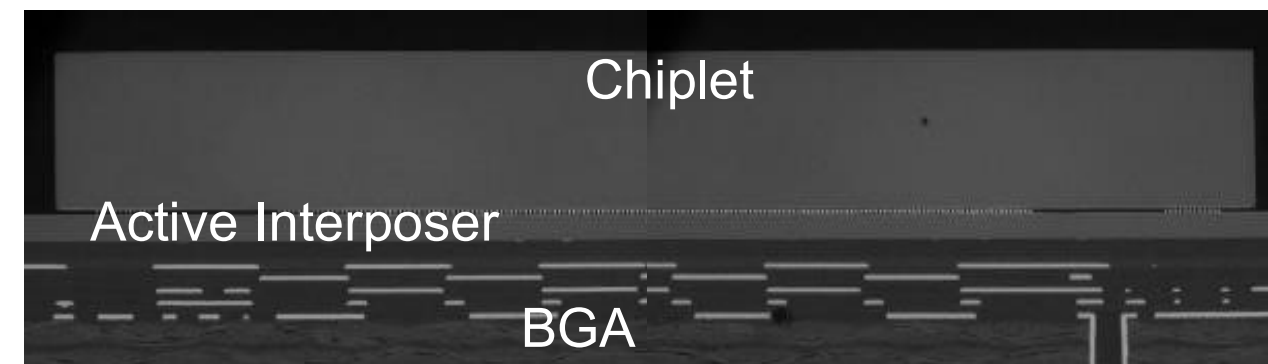


*INTACT prototype  
with open lid*



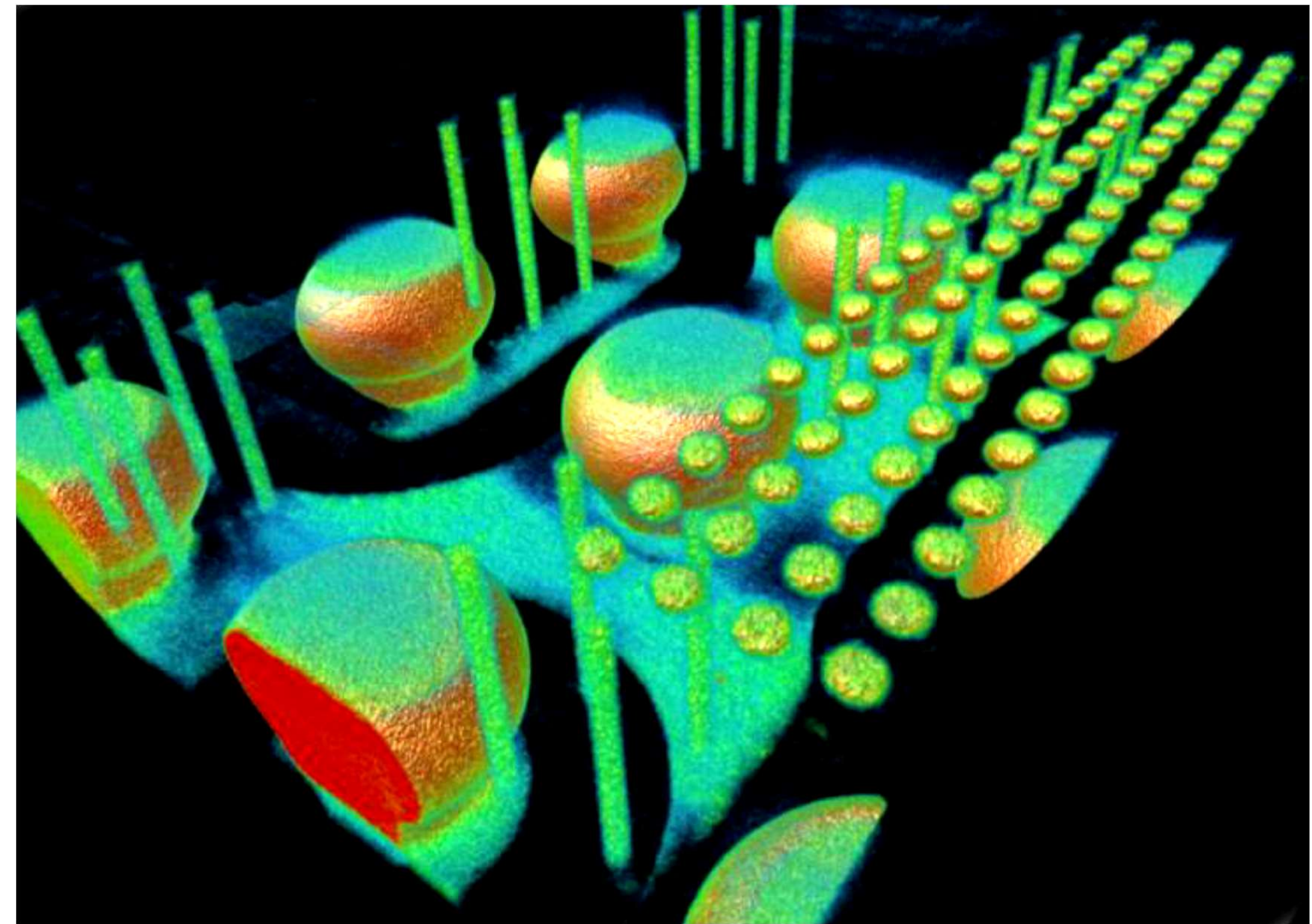
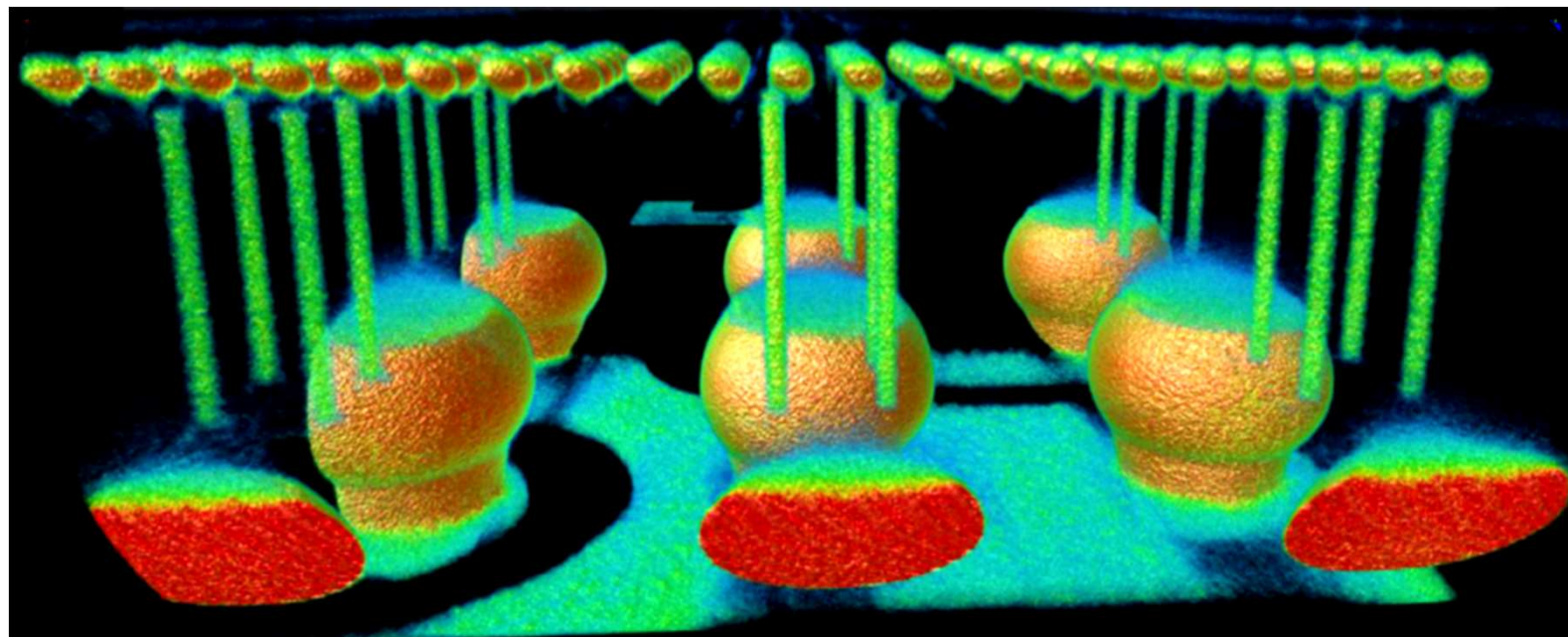
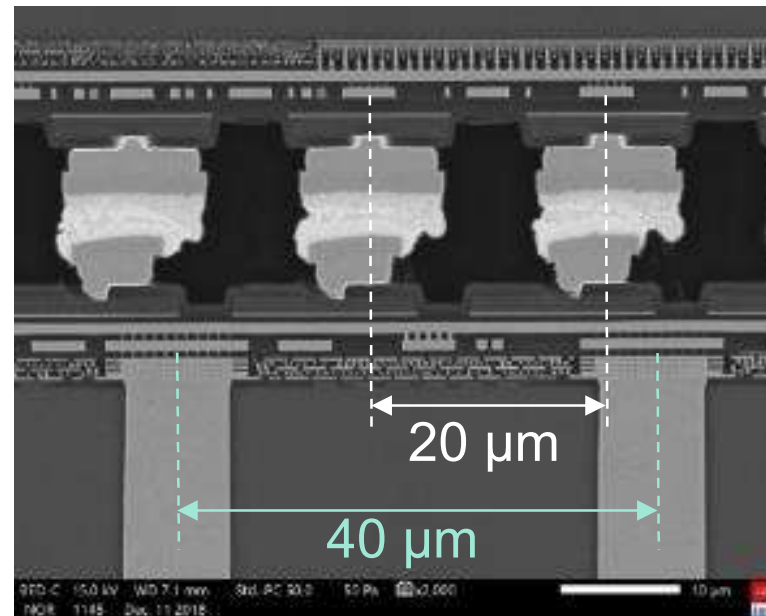
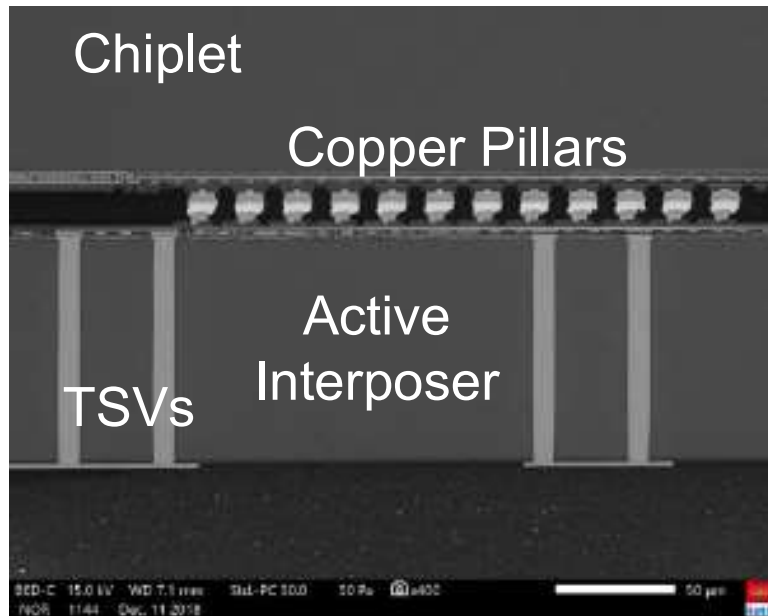
*SEM cross sections of the package  
and focus on the 3D stack*

- First demonstration of a large-scale energy efficient computing system on active interposer with chiplets 96-core architecture





# 3D assembly physical characterization



3D X-Ray tomography revealing the internal interconnects structure:  $\mu$ bumps / TSV / Bumps

# INTACT circuit performances

- **Chiplet main performances**

- Frequency: 130 MHz @ 0.5V – **1.15GHz @ 1.1V with back-bias**
- Peak performance: **220 GOPS for all 96 cores @ 1.15 GHz**
- Energy efficiency: **9.6 GOPS/W (Coremark) @ 246MHz @ 0.6V**

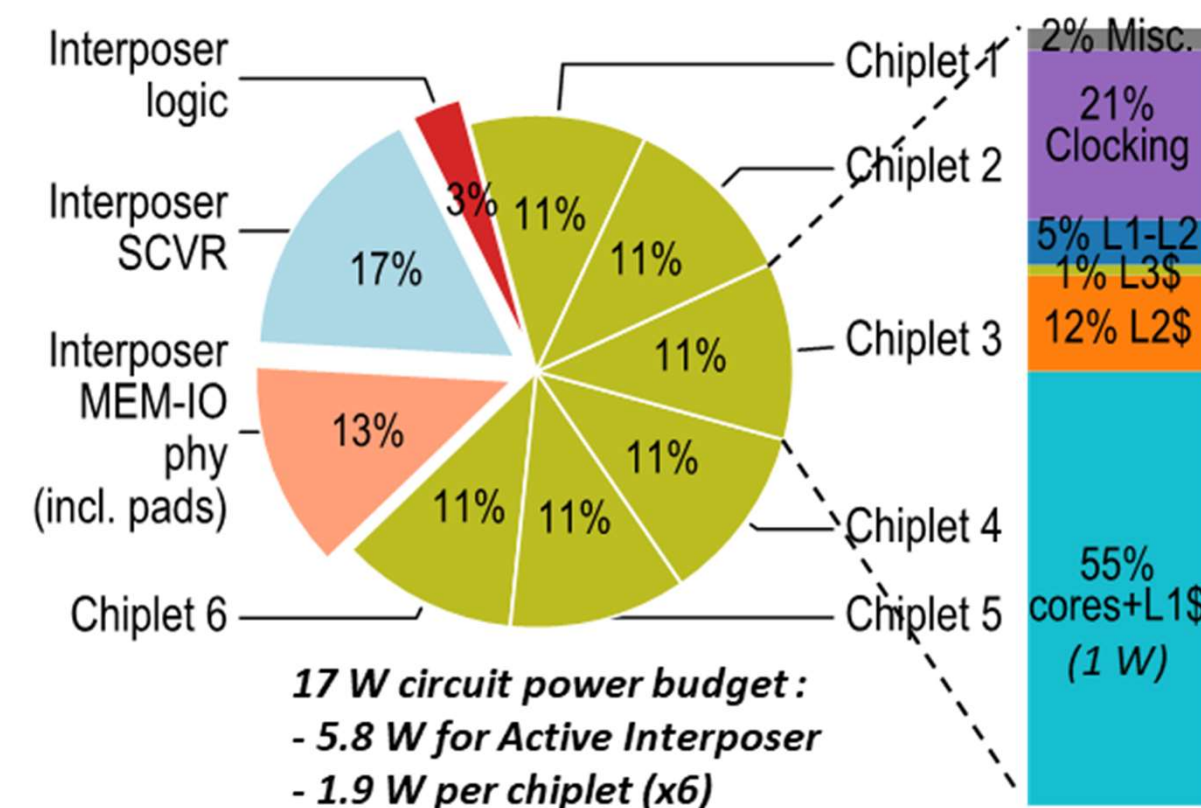


- **3D-plug interface & low latency ANOC**

- Throughput up to **3 Tbit/s/mm<sup>2</sup>**
- Latency down to **0.6 ns/mm**

- **Power consumption break-down**

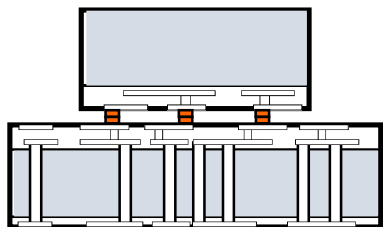
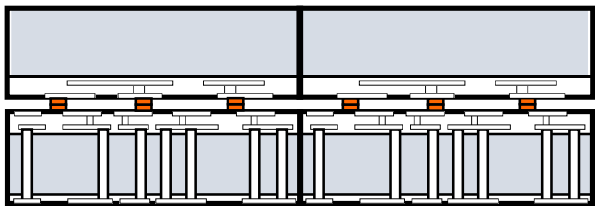
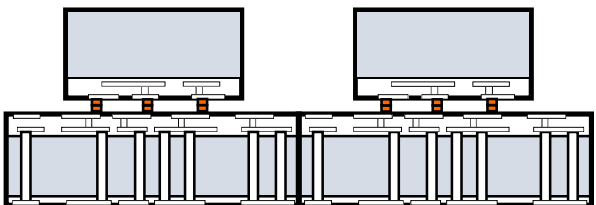
- Cores+L1: ~50% power per chiplet
- Interposer logic & interconnects : 3%
- Voltage Regulators : 17% of overall power budget



Power consumption break-down



# Challenges for inter-die fine pitch interconnects

	 Die-to-die	 Wafer-to-wafer	 Die-to-wafer
Design flexibility	OK	KO	OK
Known Good Die	OK	KO	OK
Multi-die stacking	OK	KO	OK
Fine pitch enabler	Warpage control	Hybrid bonding	Hybrid bonding
Throughput enabler	No	Natively collective	Coll. self assembly

- Pick & Place faces strong alignment / throughput tradeoff
- Die-to-wafer hybrid bonding achieves finer pitch
- Self-assembly as a throughput enabler for die-to-wafer hybrid bonding



# Direct hybrid bonding principle

- Direct bonding is based on the spontaneous adhesion of smooth surfaces
- Cu/SiO<sub>2</sub> Hybrid bonding achieves simultaneous stacking & interconnection

## Advantages

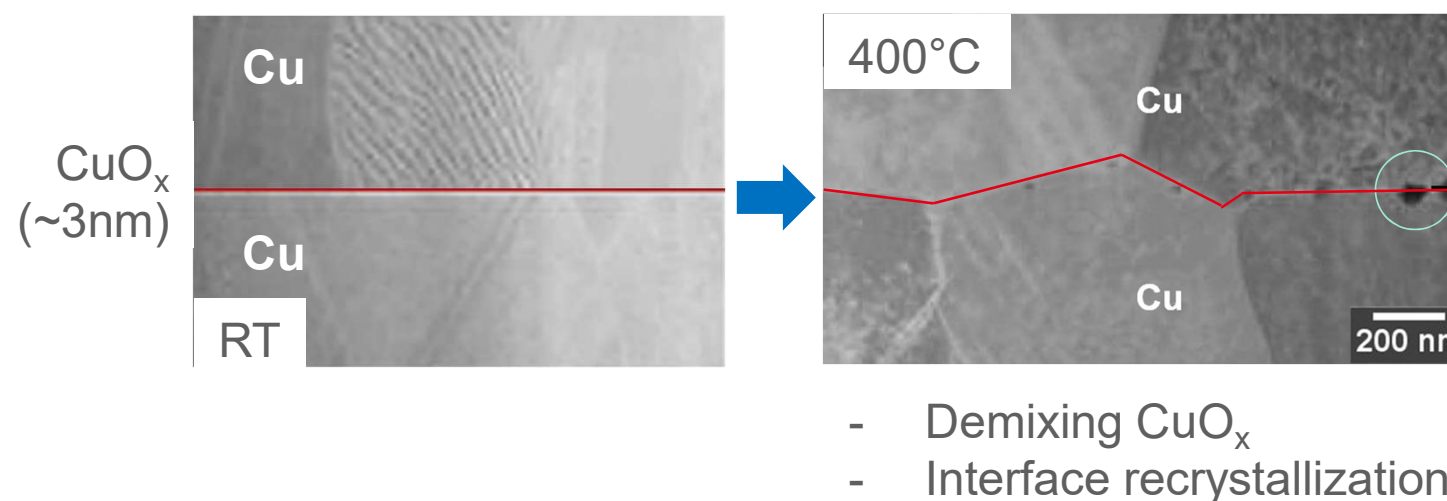
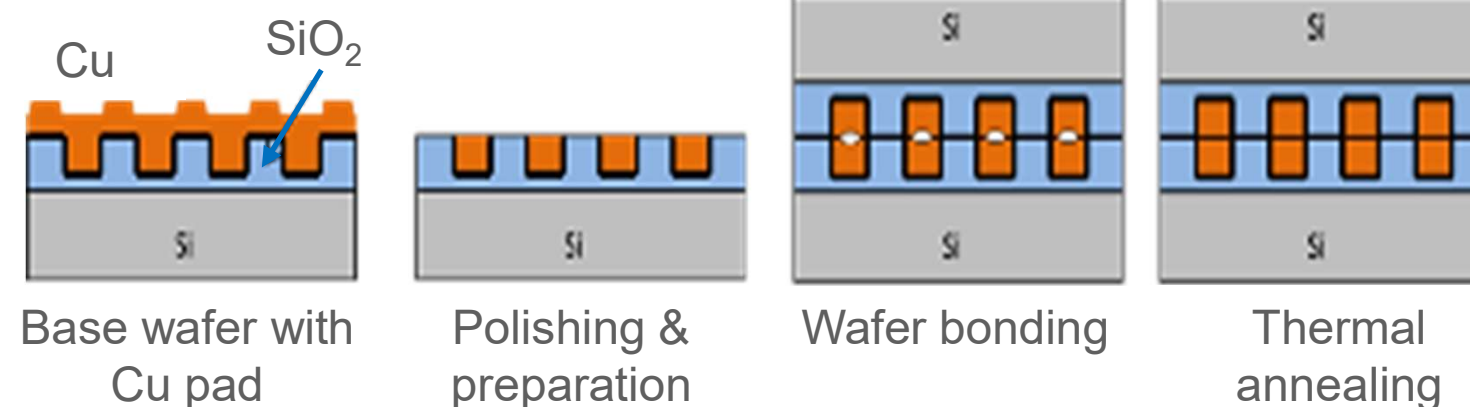
Pitch down to <1μm

No underfill, no gap

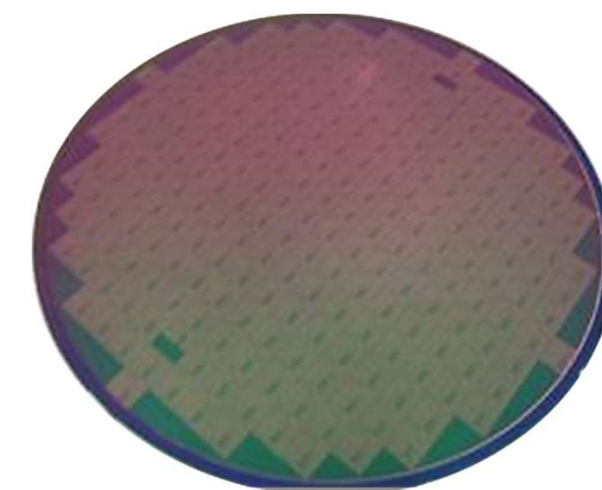
Bonding @ RT  
no compression

Wafer-to-wafer  
Or Die-to-wafer

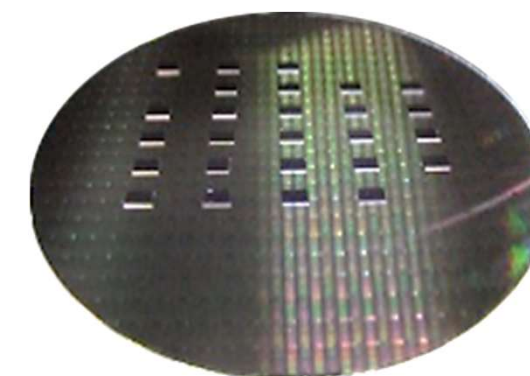
Classical  
damascene process



## Wafer-to-Wafer

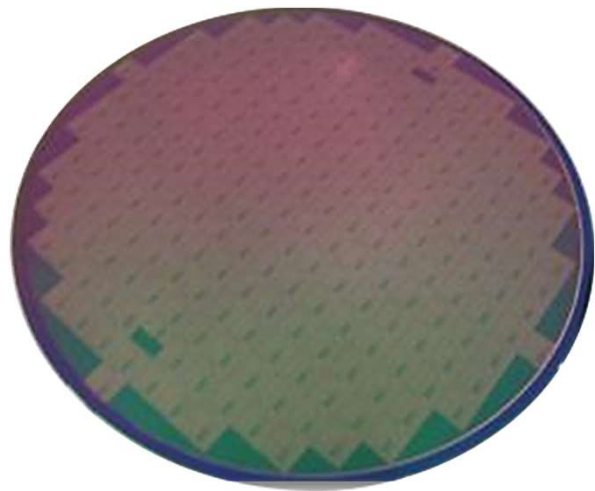


## Chip-to-Wafer

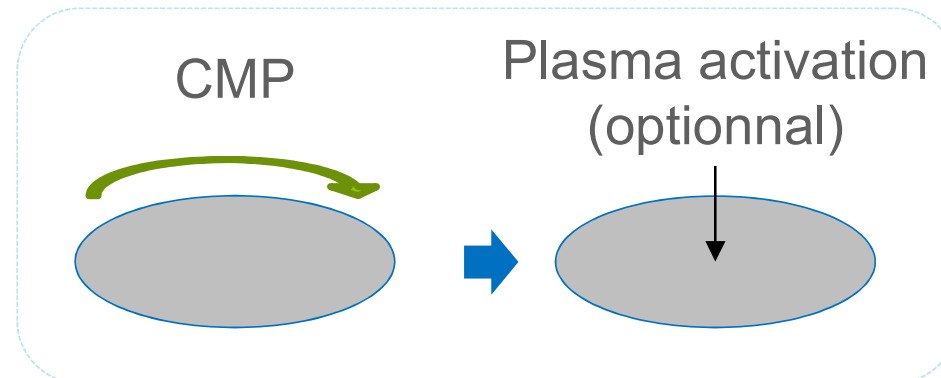
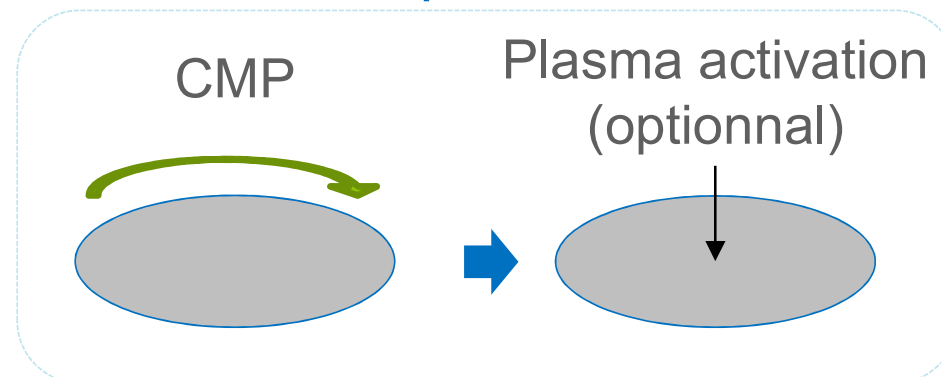


# From wafer-to-wafer hybrid bonding...

## Wafer-to-Wafer

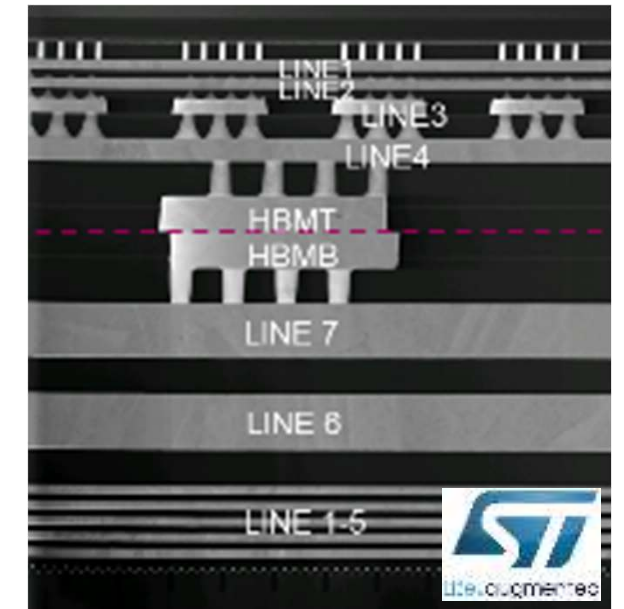
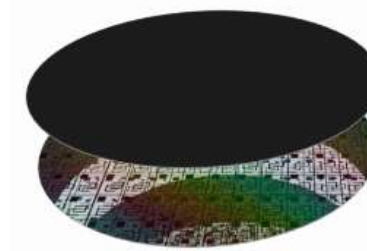


Top wafer



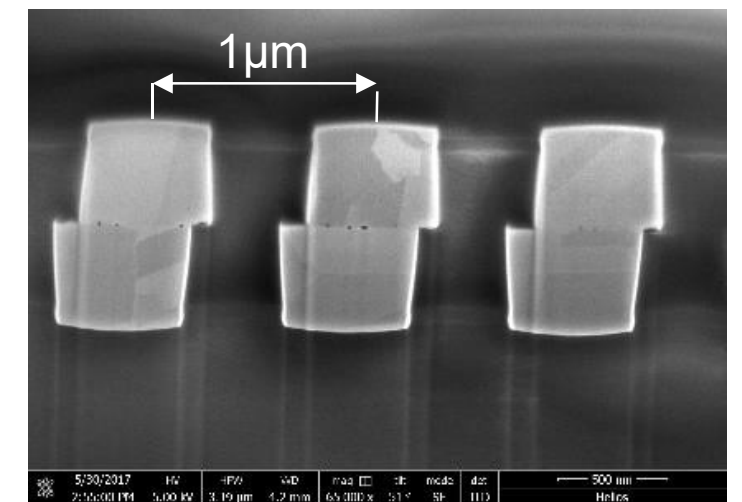
Bottom wafer

NANO ELEC.



Hybrid bonding on CMOS wafers

- Wafer-to-wafer hybrid bonding achieves **ultra fine <1μm 3D interconnects pitch** with **high throughput**
- Design limited (identical footprint between top/bottom)

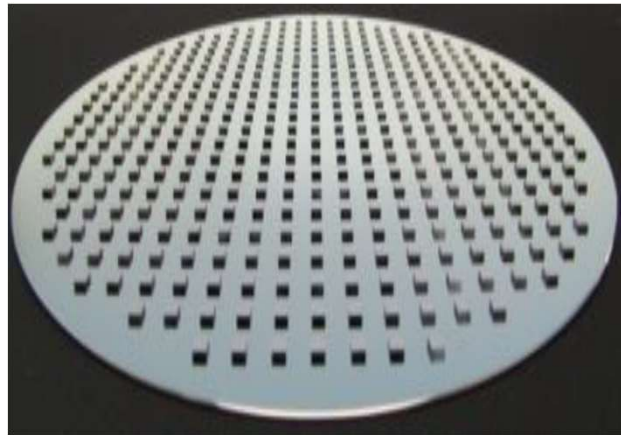


Hybrid bonded 1μm pitch interconnects

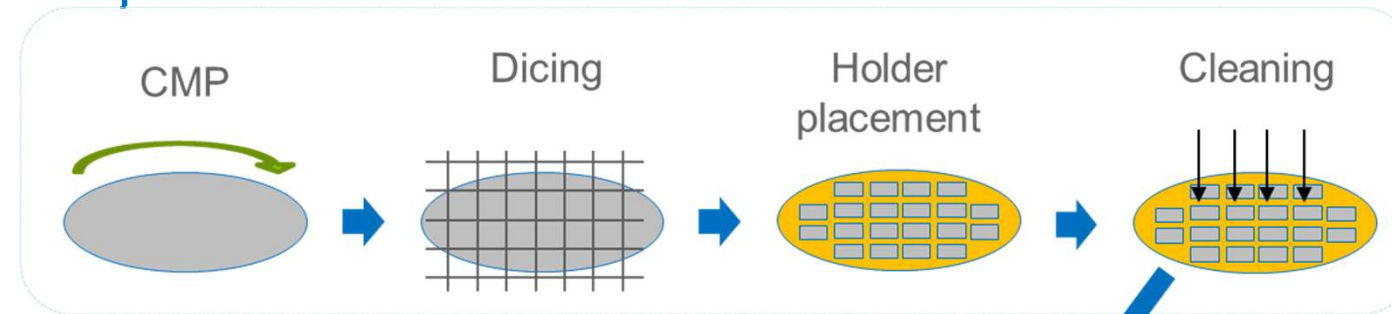


# ....to die-to-wafer hybrid bonding

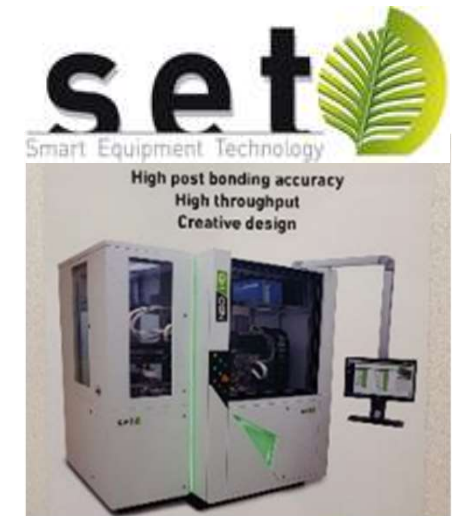
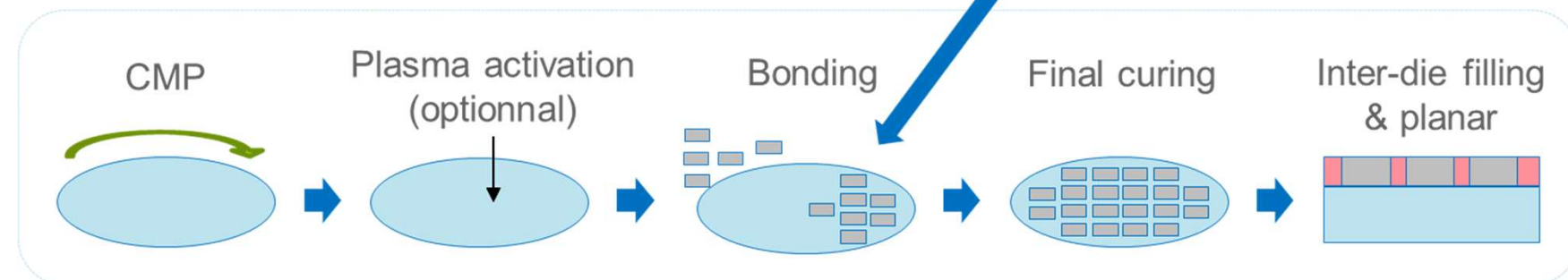
## Die-to-Wafer



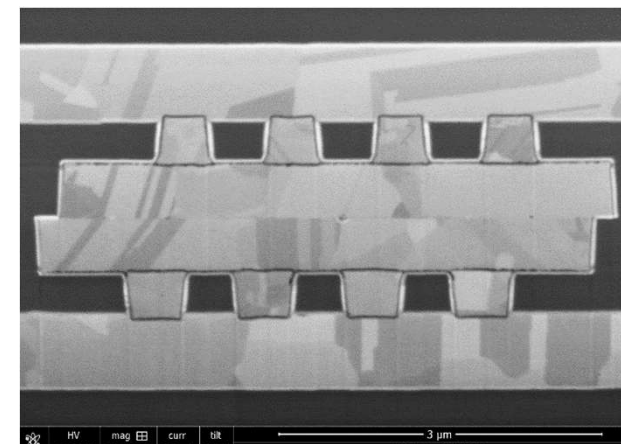
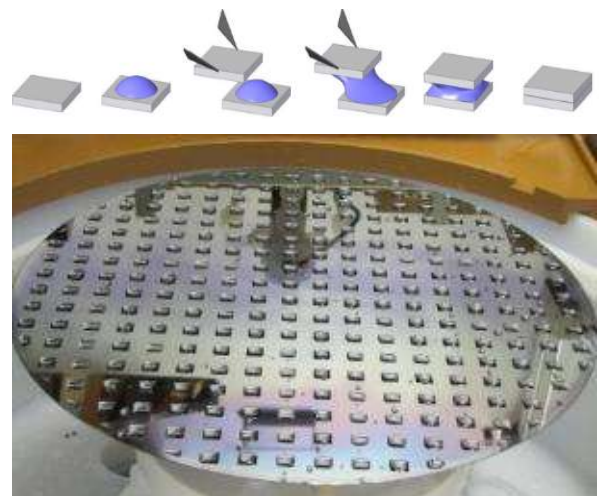
Top die



Bottom wafer



Die-to-wafer hybrid bonding SET tool



SEM cross section of die-to-wafer hybrid bonded interconnects

Self-assembly demonstration for high throughput / high accuracy

- Die-to-wafer allows **known good die** (KGD) strategy, **heterogeneity** & **multi-die** stacking
- Throughput improved with self-assembly collective approach



# Take away messages

- **Heterogeneity & bandwidth enablers for exascale HPC**
  - Chiplets on active interposer as new paradigm for HPC
  - 3D assembly options towards heterogeneous processors
  - Higher bandwidth comes with fine pitch 3D interconnects
  - Embedded power management allows energy efficiency
- **Successful chiplets integration on active silicon interposer demonstrated with high performance**
- **Further improvements underway with process developments**
  - Direct hybrid bonding for ultra-dense die-to-die interconnects
  - Die-to-wafer bonding for heterogeneous 3D integration (III/V on CMOS...)



# Thanks for your attention



This work was supported by the French National Program “Programme d’Investissements d’Avenir, IRT Nanoelec” under Grant ANR-10-AIRT-05.