2.3 A 220GOPS 96-Core Processor with 6 Chiplets 3D-Stacked on an Active Interposer Offering 0.6ns/mm Latency, 3Tb/s/mm² Inter-Chiplet Interconnects and 156mW/mm² @ 82%-Peak-Efficiency DC-DC Converters

Pascal Vivet[1], Eric Guthmuller[1], Yvain Thonnart[1], Gael Pillonnet[1], Guillaume Moritz[1], Ivan Miro-Panadès[1], Cesar Fuguet[1], Jean Durupt[1], Christian Bernard[1], Didier Varreau[1], Julian Pontes[1], Sebastien Thuries[1], David Coriat[1], Michel Harrand[1], Denis Dutoit[1], Didier Lattard[1], Lucile Arnaud[1], Jean Charbonnier[1], Perceval Coudrain[1], Arnaud Garnier[1], Frederic Berger[1], Alain Gueugnot[1], Alain Greiner[2], Quentin Meunier[2], Alexis Farcy[3], Alexandre Arriordaz[4], Severine Cheramy[1], Fabien Clermidy[1]

[1]CEA-LETI-MINATEC, Grenoble, France, [2]Sorbonne University, Paris, France, [3]STMicroelectronics, Crolles, France, [4]Mentor, St. Ismier, France

In the context of high-performance computing and big-data applications, the quest for performance requires modular, scalable, energy-efficient, low-cost manycore systems. Partitioning the system into multiple chiplets 3D-stacked onto large-scale interposers – organic substrate [1], 2.5D passive interposer [2] or silicon bridge [3] – leads to large modular architectures and cost reductions in advanced technologies by the Known Good Die (KGD) strategy and yield management. However, these approaches lack flexible efficient long-distance communications, smooth integration of heterogeneous chiplets, and easy integration of less-scalable analog functions, such as power management [4] and system IOs. To tackle these issues, this paper presents an *active* interposer integrating: i) a Switched Capacitor Voltage Regulator (SCVR) for on-chip power management; ii) flexible system interconnect topologies between all chiplets for scalable cache coherency support; iii) energy-efficient 3D-plugs for dense inter-layer communication; iv) a memory-IO controller and PHY for socket communication. The chip (Fig. 2.3.7) integrates 96 cores in 6 chiplets in 28nm FDSOI CMOS, 3D-stacked in a face-to-face configuration using 20μm-pitch micro-bumps (μ-bumps) onto a 200 mm² active interposer with 40μm-pitch Through Silicon Via (TSV) middle in a 65nm technology node. Even though complex functions are integrated, active-interposer yield is high thanks to the mature 65nm node and a reduced complexity (0.08transistors/μm²), with 30% of interposer area devoted to a SCVR variability-tolerant capacitors scheme.

As shown in Fig. 2.3.1, each chiplet is composed of 4 clusters of 4 scalar MIPS32v1 cores [5]. Memory is physically distributed throughout chiplet L2-caches, and provides a fully cache-coherent hierarchy composed of: i) 16kB L1 I-cache and D-cache per core; ii) a shared distributed L2-cache with 256kB per cluster; iii) an adaptive L3-cache with 4 tiles of 1MB per chiplet, for a total of 34MB of cache. Flexible system communications are fully distributed in the 6 chiplets *and in the active interposer* using Network-on-Chip (NoC) routers and different kind of 3D-plugs: i) between L1 and L2 caches: a 5-channel 2D-mesh interconnect implements the cache-coherency protocol, using short-reach high-bandwidth inter-chiplet passive links on-interposer; ii) between L2 and L3 tiles: a 2-channel 2D-mesh interconnect uses long-reach low-latency Quasi-Delay-Insensitive (QDI) asynchronous active links in the interposer; and, iii*)* between L3-caches and off-chip DRAM memory: a 2-channel 2D-mesh interconnect uses long-reach synchronous active links in the interposer and integrates a memory-IO controller and a 4×32b LVDS PHY at 600Mb/s offering a total 19.2GB/s off-chip bandwidth. The 3D flexible communications are allowed by generic chiplet-interposer interfaces, so-called 3D-plugs, which integrate the micro-bump array, the micro-buffer cells (bidirectional driver with ESD protection, level shifter and pull-up), and boundary-scan logic. Chiplet-level Dynamic Voltage and Frequency Scaling (DVFS) is enabled by SCVRs on the active interposer below each chiplet, surrounded by the pipelined NoC links. Finally, the active interposer embeds regular infrastructure, such as clocking, configuration, sensors (thermal & stress), and a design-for-test scheme. The circuit is fully testable (Fig. 2.3.8), for KGD sorting of both chiplets and the active interposer, and for final test, using compressed full scan, IJTAG for 2D IO and 3D IO boundary scan, BIST engines, and dedicated test pads.

The active interposer allows the integration of 6 SCVRs for individual and fast DVFS transitions per chiplet and reduced IR-drop effects. Below each chiplet, each SCVR (Fig. 2.3.2) is composed of 270 0.2×0.2μm² regular unit cells in a checkerboard pattern using thick-oxide transistors and a MOS-MOM-MIM capacitor stack to maximize the capacitance density (8.9nF/mm²). The 11.3mm² SCVR footprint, corresponding to 50% of the above chiplet's area, achieves a measured 156mW/mm² power density at 82% peak efficiency. The SCVR input voltage V_{IN} (up to 2.5V) is delivered from the interposer back-face through a 40μm-pitch TSV array, and then stepped-down by a 10-phase interleaved 3-stage gearbox scheme to generate 7 lossless voltage ratios from 4:1 to 4:3. This allows high conversion efficiency over a wide V_{OUT} range of 0.35V-1.3V for flexible DVFS, which is finally delivered to the chiplet through a micro-bump face-to-face power grid (Fig. 2.3.7). The SCVR input voltage (up to 2.5V) reduces total input current and the required number of power IOs in the package. Each SCVR is supervised by a central clock-frequency and feedback controller to allow fast DVFS and IR-drop mitigation (<10ns step response).

Figure 2.3.3 illustrates the microarchitecture of the source-synchronous 3D-plugs used for 2.5D passive and 3D face-to-face links. Implemented as a standard synthesizable digital design, 3D-plugs provide multiple Virtual Channels (VC) and use credit and clock forwarding schemes. They operate at a higher frequency than the NoC to reduce contention due to VC multiplexing. Delay lines and polarity selectors are used to skew the TX clock for RX data sampling (CLK_TX_ϕ1) and TX credit sampling (CLK_TX_ϕ2). 2.5D passive links are routed using M2-M4 or M3-M5 BEoL metals with 0.3μm width, 1.1μm pitch and balanced track lengths, while the forwarded clocks are routed separately with ground shielding. The achieved 3Tb/s/mm² bandwidth density is 1.9× higher than [2] for 3D links, and 2.5D passive links reach a 12% higher bandwidth cross-section. The aggregate synchronous 3D/2.5D links bandwidth is 527 GB/s.

The different inter-chiplet interconnects are detailed in Fig. 2.3.4. Their system-level performance is measured using on-chip traffic generators and probes, while latency and energy breakdowns have been refined by simulation. The three interconnect types show different strengths and tradeoffs, which can be best used for different traffic types. Abutted synchronous 3D-plugs for neighboring L1-L2 cache coherence traffic operate up to 1.25GHz with the lowest latency of 7.2ns between source and destination clock domains. Large applications with distant L1-L2 traffic require routing over several chiplets, with the lowest propagation energy of 0.15pJ/b/mm due to a full path in the 28nm node, but a latency increased by multiple clock-domain crossings. QDI asynchronous logic for L2-L3 traffic has the lowest latency of 0.6ns/mm including routers and pipelined links, which is 3.3× better than its synchronous counterpart, with an efficient 0.97GHz 4-phase protocol in the interposer and 2-phase conversion to hide the handshake latency in the 3D-plugs. L3-to-off-chip-memory traffic uses a single clock domain in the interposer with best synchronous propagation speed of 2ns/mm.

Figure 2.3.5 explores overall chip power and performance measurements. Power consumption and energy efficiency, while running the Coremark benchmark, is compared to a theoretical system using a digital LDO instead of the proposed fully integrated SCVR. Using a LDO at same V_{IN} = 2.5V would result in a 2× increase in power consumption, a lower V_{IN} would be needed to limit losses at the expense of more power pins and voltage-drop issues. The power breakdown shows the low power budget of the active interposer (only 3% is devoted to interposer logic). The cores represent over half the power consumption of the chiplets, consuming the majority of the measured chip power (17W). Lastly, scalability of the cache-coherent architecture is analyzed by running an image filtering application from 1 to 512 cores. Results for more than 96 cores were obtained by RTL simulation with additional chiplets. Compared to single core execution, a 67× execution-time speedup is obtained with 96 cores and 340× with 512 cores.

Compared to prior art (Fig. 2.3.6), the circuit is the first chiplet-based manycore architecture for high-performance computing using an active interposer, integrating: i) a fully integrated voltage regulator, using free area available in the active interposer, offering DVFS-per-chiplet and achieving 156mW/mm² at 82% peak power efficiency (10% higher than [4]); ii) flexible and distributed NoC meshes for scalable cache-coherency traffic, with 0.6ns/mm inter-chiplet latency using asynchronous signaling, and a 0.59pJ/b synchronous 3D-plug energy efficiency with 3Tb/s/mm² bandwidth density.

An active interposer enables efficient integration of large-scale chiplet-based computing systems. Such schemes can be applied for integration of similar chiplets as presented above, but also for smooth integration of heterogeneous chiplets.

Acknowledgements:
This work was partly funded by the French National Program Programme d'Investissements d'Avenir, IRT Nanoelec under Grant ANR-10-AIRT-05.

References:

[1] N. Beck et al., "Zeppelin: An SoC for Multichip Architectures," *ISSCC*, pp. 40-41, Feb. 2018.

[2] Mu-Shan Lin et al., "A 7nm 4GHz Arm®-Core-Based CoWoS® Chiplet Design for High Performance Computing," *IEEE Symp. VLSI Circuits*, June 2019.

[3] David Greenhill et al., "A 14nm 1GHz FPGA with 2.5D Transceiver Integration," *ISSCC*, pp. 54-55, Feb. 2017.

[4] P. Meinerzhagen et al., "An Energy-Efficient Graphics Processor Featuring Fine-Grain DVFS with Integrated Voltage Regulators, Execution-Unit Turbo, and Retentive Sleep in 14nm Tri-Gate CMOS," *ISSCC*, pp. 38-40, Feb. 2018.

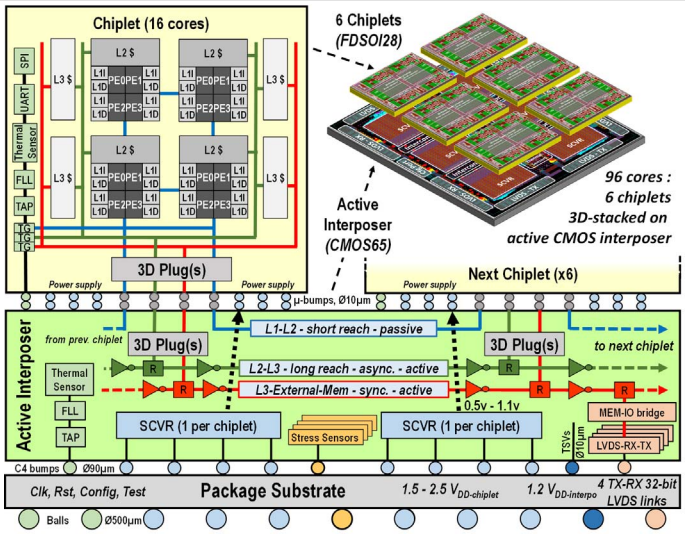


Figure 2.3.1: 96-core architecture composed of 6 chiplets 3D-stacked onto an active CMOS interposer.

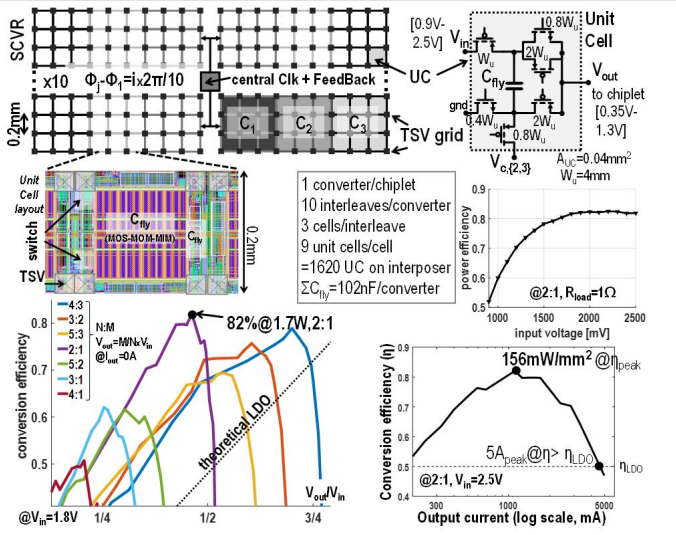
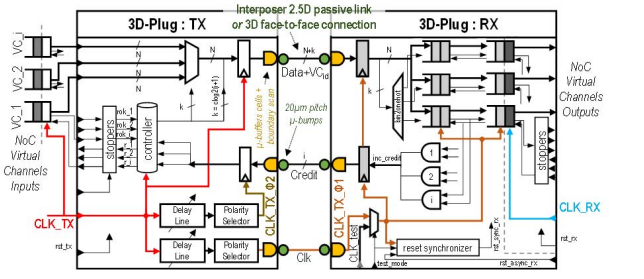
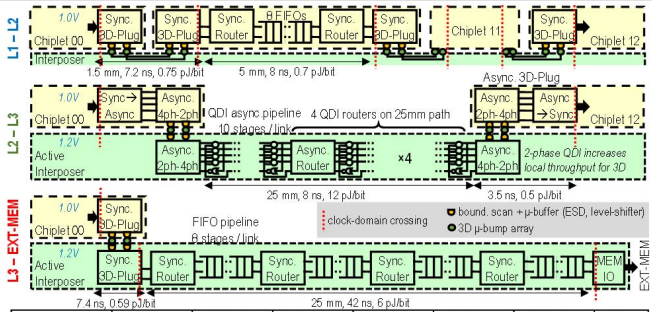


Figure 2.3.2: Switched Capacitor Voltage Regulator (SCVR) providing DVFS per-chiplet.



	This work		[2] VLSI'19	Units
Technology	28nm FDSOI chiplets / 65nm active interposer		7nm FinFet chiplet	
3D Link type and technology	2.5D	3D	LIPINCON™	
	Passive (65nm Si)	Active (face-to-face)	Passive (CoWoS™)	
System Integration	L1$ ↔ L2$	L3$ ↔ Ext. Mem	L2$ ↔ L3$	
Bus Width	168	156	320	pins
# On-Chip Bus Links	6+8+1	6	2	
Channel Length	**1.5 - 1.8**	**0.05**	0.5	mm
Die-to-Die Bump Pitch	**20**		40	µm
Voltage swing	1.2		0.3	V
Data Rate	1.25	1.21	8	Gb/s/pin
Power Efficiency	0.75	0.59	0.56	pJ/bit
Bandwidth Density	0.5	**3.0**	1.6	Tb/s/mm²
Bandwidth Cross Section	**0.9**	N/A	0.8	Tb/s/mm
Aggregate inter-Chiplets Bandwidth	394	133	640	GB/s
	527			

Figure 2.3.3: 3D-plug communication: source-synchronous interface microarchitecture and performance.



	L1-L2 nearest	L1-L2 farthest	L2-L3 4-phase	L2-L3 2-phase	L3-EXT-MEM	3DNOC [6]	Units
Reach	1.5	15	25 (bottom left chiplet to upper right chiplet)			8*	mm
Word size	40		72			32	bits
Interposer	1 passive link	3 passive links	Active async. routing	Active async. routing	Active sync. routing	Active async. Routing	—
Chiplet	—	Global sync. routing	Local sync. routing	Local sync. routing	Local sync. routing		—
3D Plug frequency	**1.25**	1.25	0.30	**0.52**	1.21	0.32	GHz
2D NoC frequency	—	**1.00**	**0.97**		0.75	0.89	GHz
End to end latency	2x4+[0-1]	44	4 + async.	4 + async.	37	4 + async.	cycles
	7.2	44.0	**15.2**	**15.2**	49.5	10*	ns
Propagation speed	4.8	2.9	**0.6**	**0.6**	**2.0**	1.2*	ns/mm
Energy / bit / mm	0.29	**0.15**	0.52	0.52	0.24	0.5*	pJ/bit/mm

(* estimated from published data)

Figure 2.3.4: Interposer distributed synchronous and asynchronous NoCs; 3D-plug asynchronous interface.

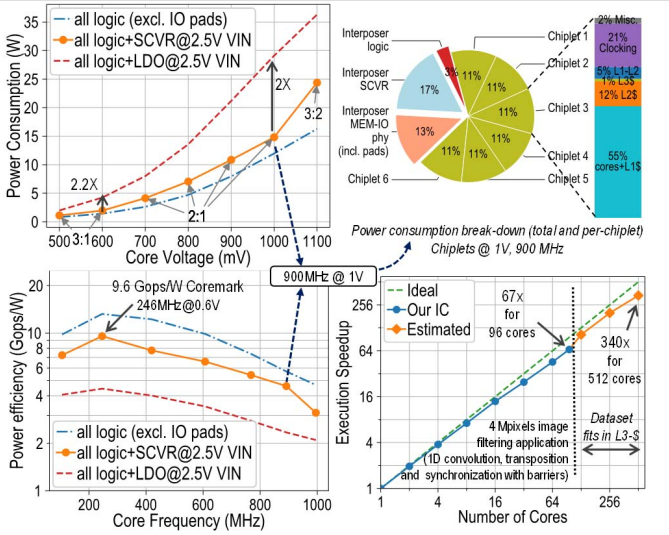
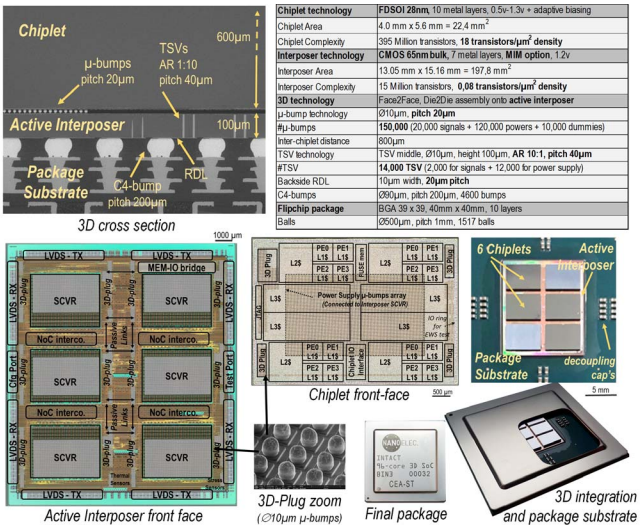


Figure 2.3.5: System performance, including SCVR<->LDO comparison, power breakdown, and many-core scalability.

		This work	[4] ISSCC'2018 INTEL	[1] ISSCC'2018 AMD	[2] VLSI'2019 TSMC	[3] ISSCC'2017 INTEL	Units
Technology	Chiplet Technology	FDSOI 28nm	FinFET 14nm	FinFET 14nm	FinFET 7nm	FinFET 14nm	
	Interposer Technology	**Active** CMOS 65nm	no	MCM substrate	Passive CoWoS ®	EMIB bridge	
	Interposer extra features	**yes**	*N/A*	no	no	no	
	Total system yield	High, using active interposer mature technology and low transistor count	*N/A*	high	high	high	
	Die-to-Die µbump pitch	**20**	*N/A*	> 100	40	55	*µm*
Power Mgt	Voltage Regulator (VR) type	**Integrated in interposer,** 1 SCVR per chiplet with **MOS+MOM+MIM**	on-chip distributed SCVR with MIM	LDO per core, with MIM	no	no	
	VR area	34% of active interposer	MIM above 40% of core area	-	*N/A*	*N/A*	
	VR peak efficiency	**82%**	72%	LDO limited	*N/A*	*N/A*	
Interconnect	Interconnect types	**Distributed NoC meshes** for scalable chip-to-chip cache-coherency traffic	*N/A*	Scalable Data Fabric (SDF)	LIPINCON™ links	AIB interconnect	
	3D Plug power efficiency	0.59	*N/A*	2.0	0.56	1.2	*pJ/bit*
	BW density	**3.0**	*N/A*	-	1.6	1.5	*Tb/s/mm²*
	Aggregate 3D bandwidth	527	*N/A*	-	640	504	*GByte/s*
CPU	Number of chiplets	**6**	1	1 - 4	2	1 FPGA fabric 6 transceivers	
	Number of cores	**96**	18	8 - 32	8	FPGA fabric	
	Max Frequency	1.15	0.4	4.1	4	1	*GHz*
	Gops (32b-Integer)	**220 (peak mult./acc.)**	14.4	131,2 - 524,8	128	*N/A*	*Gop/s*

Figure 2.3.6: Performance summary and comparison to prior art.



Chiplet technology	**FDSOI 28nm**, 10 metal layers, 0.5v-1.3v + adaptive biasing
Chiplet Area	4.0 mm x 5.6 mm = 22,4 mm^2
Chiplet Complexity	395 Million transistors, **18 transistors/µm^2 density**
Interposer technology	**CMOS 65nm bulk**, 7 metal layers, **MIM option**, 1.2v
Interposer Area	13.05 mm x 15.16 mm = 197,8 mm^2
Interposer Complexity	15 Million transistors, **0,08 transistors/µm^2 density**
3D technology	Face2Face, Die2Die assembly onto **active interposer**
µ-bump technology	Ø10µm, **pitch 20µm**
#µ-bumps	**150,000** (20,000 signals + 120,000 powers + 10,000 dummies)
Inter-chiplet distance	800µm
TSV technology	TSV middle, Ø10µm, height 100µm, **AR 10:1, pitch 40µm**
#TSV	**14,000 TSV** (2,000 for signals + 12,000 for power supply)
Backside RDL	10µm width, **20µm pitch**
C4-bumps	Ø90µm, pitch 200µm, 4600 bumps
Flipchip package	BGA 39 x 39, 40mm x 40mm, 10 layers
Balls	Ø500µm, pitch 1mm, 1517 balls

Figure 2.3.7: Chip microphotographs, 3D cross section, package and technology features.

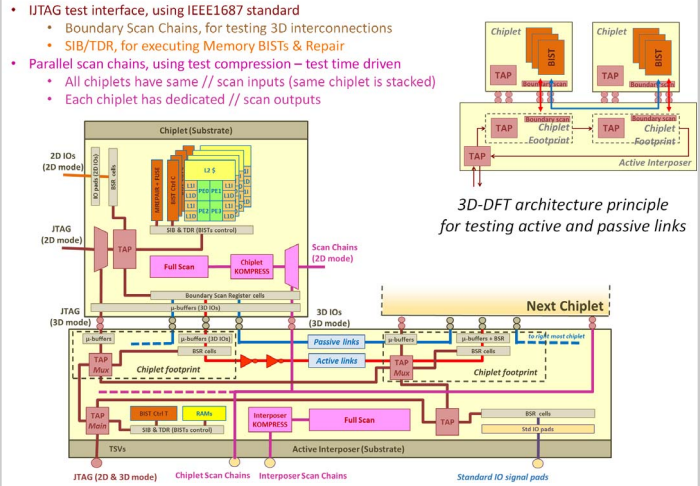


Figure 2.3.S1: 3D design-for-test architecture for Known-Good-Die (KGD) sorting (chiplet test at wafer level before assembly) and final test (final chip test after 3D assembly of chiplets onto the active interposer and packaging).

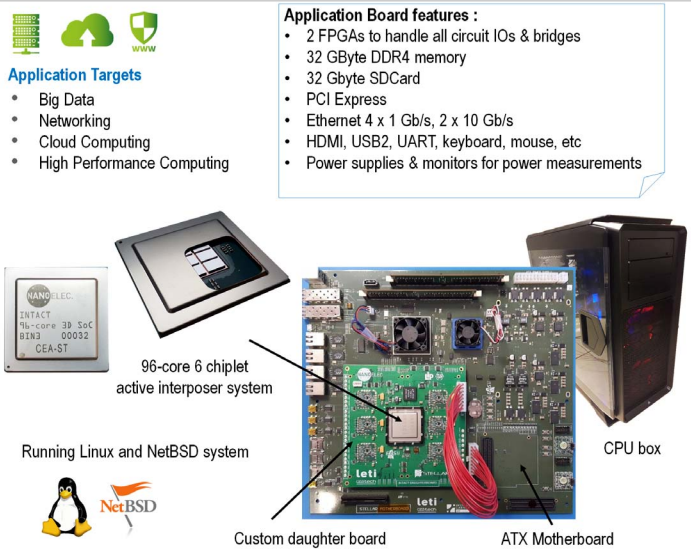


Figure 2.3.S2: Circuit board demonstration and applications.

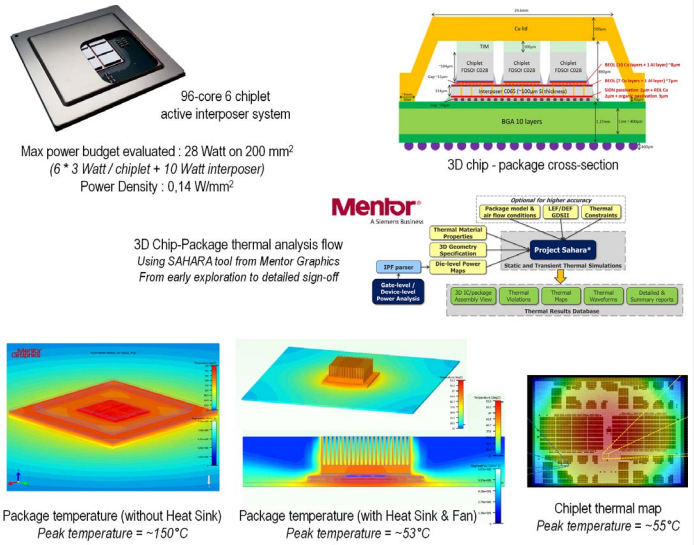


Figure 2.3.S3: 3D chip-package thermal analysis.

Additional References:
[5] E. Guthmuller et al., "A 29 Gops/Watt 3D-Ready 16-Core Computing Fabric with Scalable Cache Coherent Architecture Using Distributed L2 and Adaptive L3 Caches," *ESSCIRC*, Sept. 2018.
[6] P. Vivet et al., "A 4×4×2 Homogeneous Scalable 3D Network-on-Chip Circuit with 326 MFlit/s 0.66pJ/b Robust and Fault Tolerant Asynchronous 3D Links," *ISSCC*, pp. 146-147, Feb. 2016.

978-1-7281-3205-1/20/$31.00 ©2020 IEEE