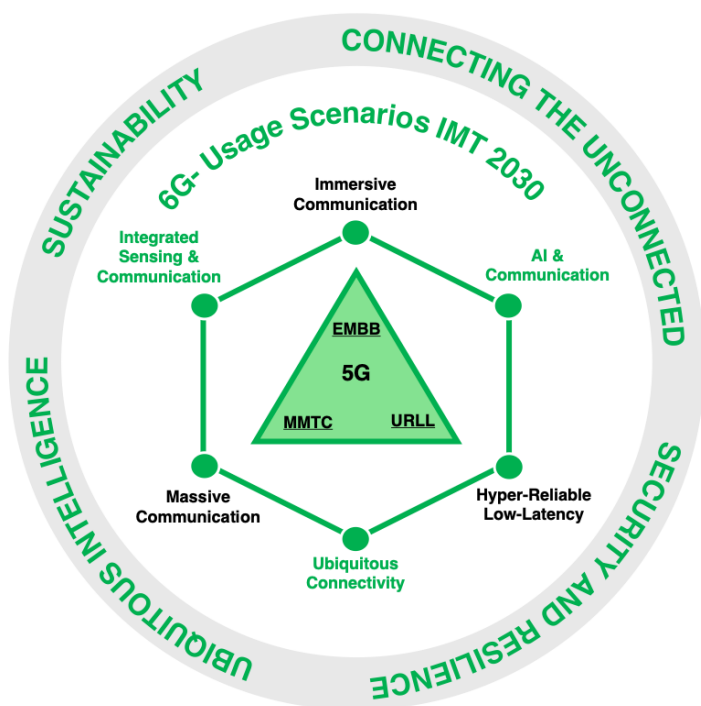


A 3D-Stacked 8.4TFLOPS/4.3W Many-Core Domain-Specific Processor for AI-Native Radio Access Networks

Marco Bertuletti¹, Yichao Zhang¹, Diyou Shen¹, Alessandro Vanelli-Coralli¹², Frank K. Gürkaynak¹, Luca Benini¹²
¹IIS, ETH Zurich; ²University of Bologna;



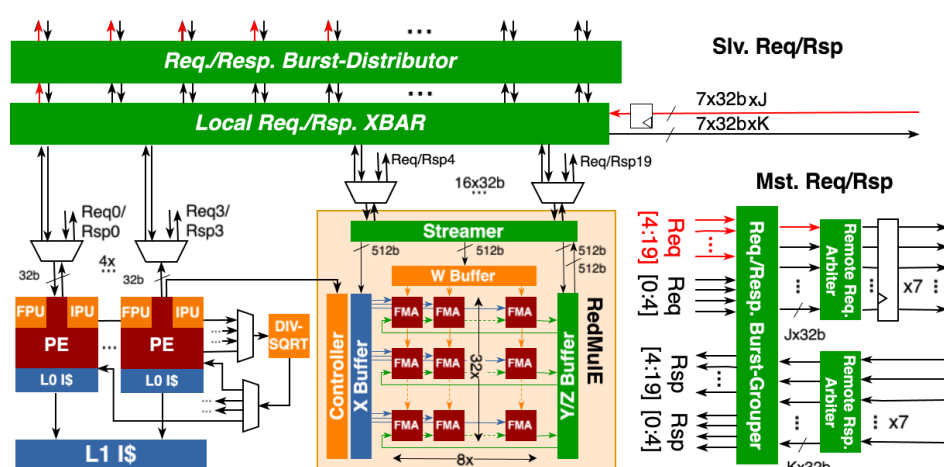
6G is coming fast!

20Gbps → 1Tbps **50x** more throughput

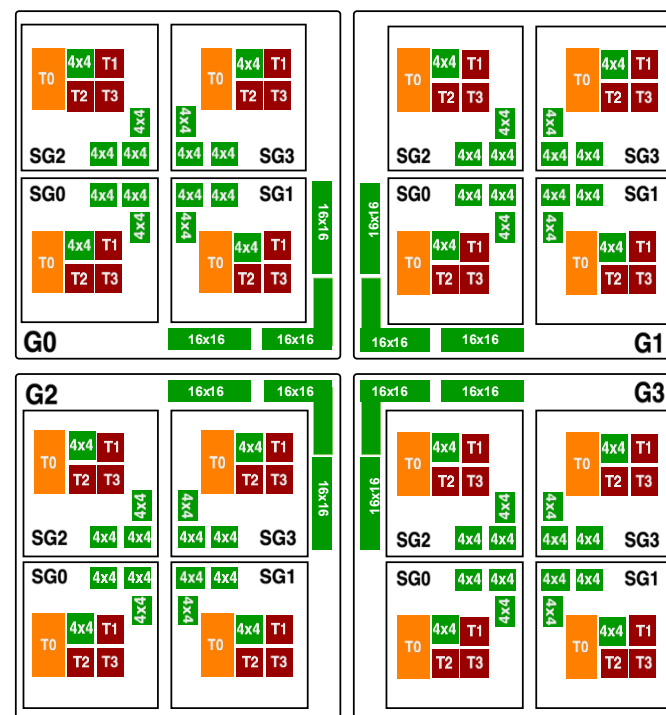
Many new compute-hungry use-cases (AI-Native RAN)

We need: **software programmability** + more **performance** + more **energy-efficiency**

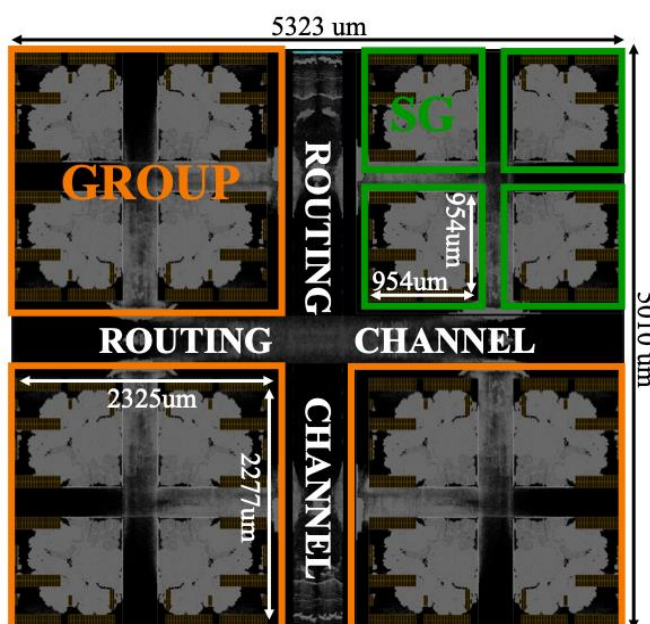
Domain specialization + 3D integration to increase compute density is needed!



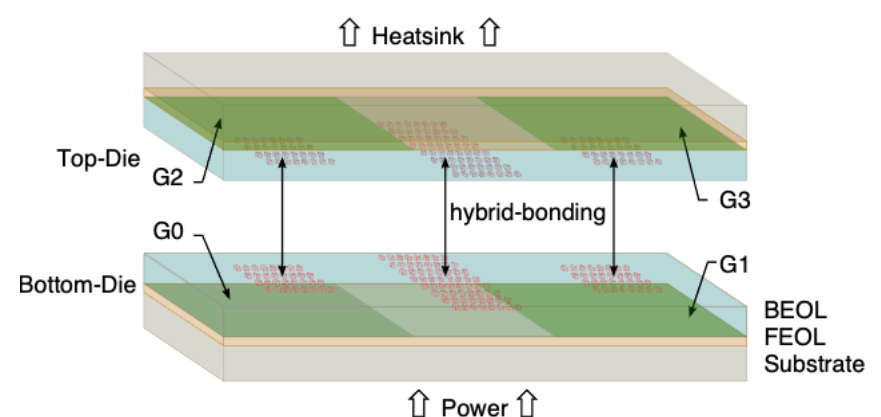
Heterogeneous Tiles with tensor engines (TEs) & flexible instruction processors (PEs)



Hierarchical low-latency Interconnect
 256 PEs, 16 TEs, shared memory for high data-reuse =
4096 MAC/cycle



89% utilization on parallel GEMM 2D N7
Implementation: 6x more throughput, **9.1x** more energy-efficiency than homogeneous baseline



2.3x super-linear footprint improvement with 3D integration.