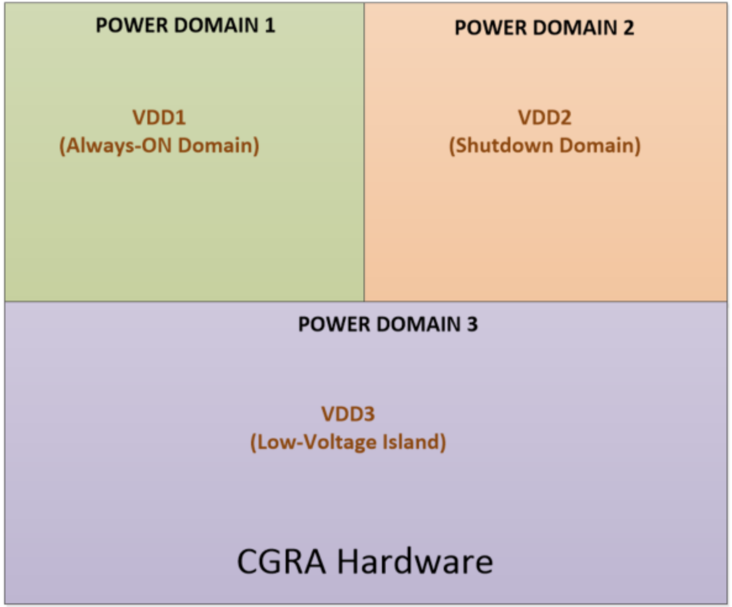
Next-Gen CGRA Architecture

System Architecture (1/2)

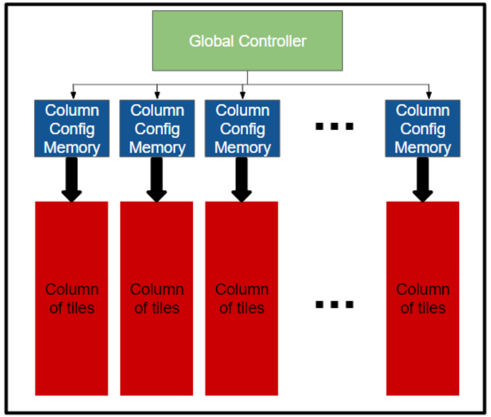
- Clock domain based CGRA
  - Allow different part of the chips to operate at different frequencies
- Power domain based CGRA
  - Multiple power domains on the chip – Low-voltage, Always-on, Shut-down
- DVFS (Dynamic voltage frequency scaling) support
  - Variable operating voltage & frequency. DVFS can be used for non-timing critical applications



Power Domain based CGRA

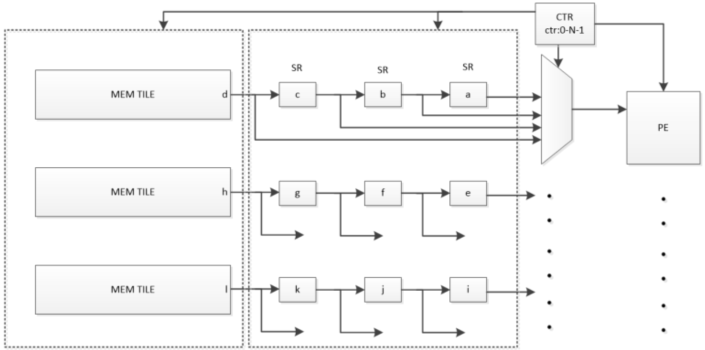
System Architecture (2/2)

- Fast reconfiguration/Run-time reconfigurability
  - Configuration currently takes ~100 JTAG clock cycles per configuration register
  - Ability to swap out applications quickly
  - E.g. mapping different convolution layers in CNN



Representative fast-reconfiguration architecture

- CGRA Hardware Virtualization
  - Resource re-use when application size is bigger than available hardware resources
  - Multi-rate support for different applications
  - State machine based design



Resource sharing in CGRA

PE Architecture

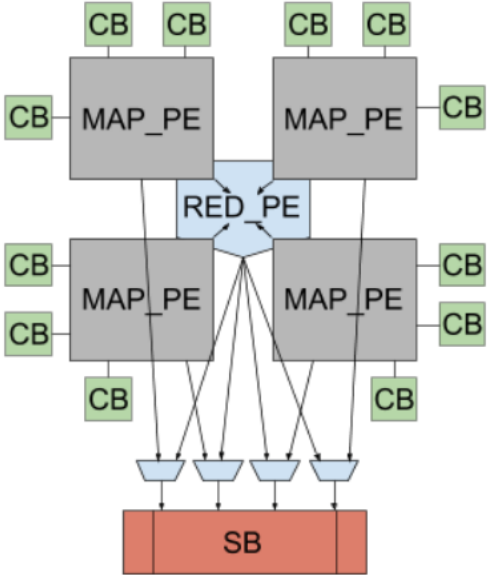
- Analyze the target applications to find frequently occurring common patterns/compute kernels

- Improve compute density for PEs
  - Current version uses 2 input-PE
  - 3-input based PE design
    - Absolute difference addition, FMA

- Support new operations – division, modulo, saturation etc.

- Super-PE architecture
  - Multi-precision handling
  - Virtualization

- Homogeneous vs Heterogeneous PE tiles
  - All PEs do not need to support all the functionalities – Improve compute density E.g. LUTs, register files



PE Clustering

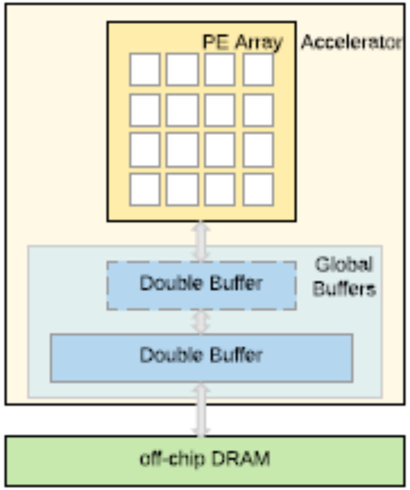
Memory Architecture

- Dual ported memory
  - Flexible and simpler design compared to single-ported SRAM and additional logic
  - Enable irregular access patterns

- Double buffering
  - Supply operands out of one buffer while performing LD/ST operations on the other
  - Allow hardware to overlap computation and memory loads

- Memory hierarchies
  - Optimize data locality

- DMA Engine
  - Mem-tile to Mem-tile copying or moving of data within memory



Architecture template with double buffer & memory hierarchy

CGRA Connectivity/Routing

- Explore dataflows/customized connectivity patterns for a class of applications
- Interconnect architecture exploration
  - Trade-offs between flexibility, fanouts and different topologies
- Routing Profile Extraction
  - List of key switchbox parameters needed for P&R to better evaluate metrics like IO utilization, wirelength length stats, routing density etc.