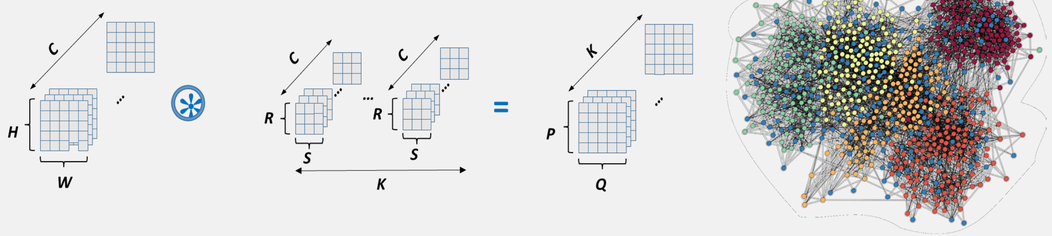
Intel SDH PUMA Platform

Fabrizio Petrini, Joshua Fryman, Matthew J. Andrus
Intel Extreme Scale Computing and Parallel Computing Laboratory

Specialized Functions: Software Defined Hardware (SDH)

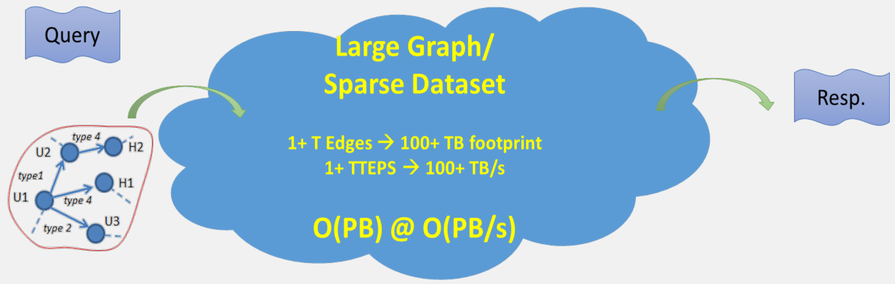
BACKGROUND

The Intel Programmable Unified Memory Architecture (PUMA) SDH proposal brings **sparse** and **dense** compute together in a transformational solution that will deliver orders of magnitude improvements in performance and energy-efficiency over conventional systems. PUMA SDH enables large-scale, real-time streaming analytics combining Deep Learning, Graph Analytics and traditional machine learning.



OVERVIEW

- PUMA adopts an application-driven approach, starting from flexibility and scalability as the primary objectives.
- PUMA attacks the structural bottlenecks that limit scalability in the network and system architecture.
- PUMA will achieve several orders-of-magnitude improvements in performance and energy efficiency.

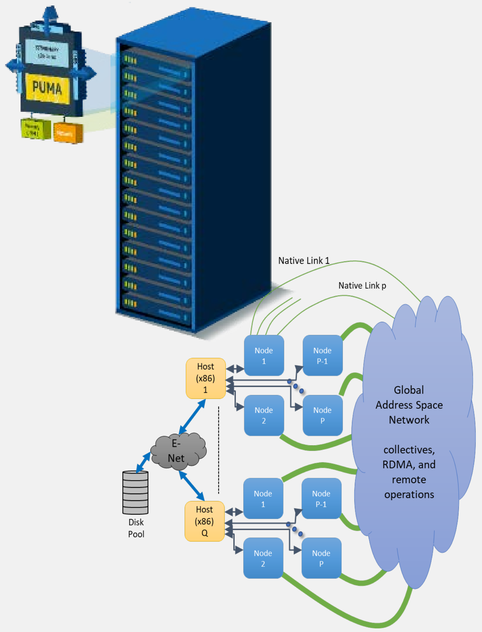


Intel's approach hinges on three dimensions

1. Early adoption of transformational technologies (memory, network, accelerators, packaging, etc.).
2. A novel system design to optimally aggregate and re-configure resources at runtime based on application demands.
3. An application driven co-design approach identifying common kernels, formalizing performance bounds and taking advantage of novel system capabilities.

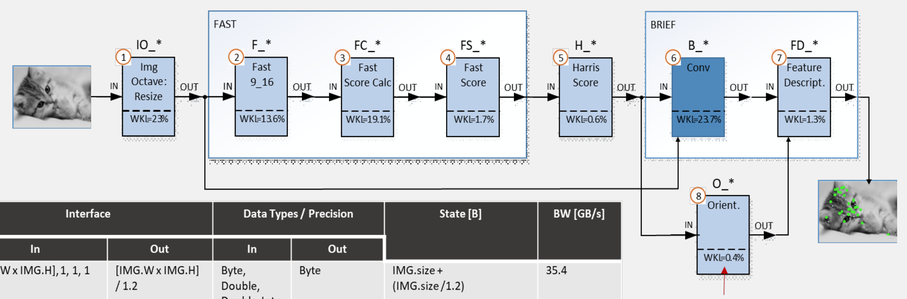
APPROACH

- We will design a new system architecture that takes advantage of a high-bandwidth fabric to enable flexible and dynamic recombination/ aggregation of resources according to workload needs.
- PUMA SDH is a flexible system that can get close to peak performance for each phase of a complex application.
- The PUMA fabric quickly reshuffles data formats and layouts between phases and performs scalable collective operations.
- PUMA SDH provides an innovative framework to seamlessly integrate sparse and dense compute, and explore novel parallelization techniques.
- HW and SW architectures are driven by the application analysis in a tight co-design



EXAMPLE RESULTS

The "Oriented FAST and Rotated BRIEF" (ORB) algorithm for Interest Point Detection incorporates several algorithms used in the image processing domain to address the key points of both feature detection and feature description problems.



ID	App Kernel	Interface		Data Types / Precision		State [B]	BW [GB/s]
		In	Out	In	Out		
1	Resize	[IMG.W x IMG.H], 1, 1, 1	[IMG.W x IMG.H] / 1.2	Byte, Double, Double, Int	Byte	IMG.size + (IMG.size /1.2)	35.4
2	Fast9_16	[16 x 1], 1	1	Byte	Int	((IMG.H+16)*3*5)+128	34.8
3	Fast Score Calc	[25 x 1], 1, 1	[CORN x 1]	Byte, Int, Int	Byte	25	20.6
4	Fast Score	[CORN x 1]	[KP x 5]	Byte	Float, Float, Float, Int, Float	CORN*10 + CORN*18	5.7
5	Harris Score	[IMG.W x IMG.H], [8 x 1], [KP x 2]	[KP x 1]	Byte, Int, Int	Float	196 + (KP.size * 8)	17.1
6	Convolution	[IMG.W+6 x IMG.H+6], [7 x 1]	[IMG.W x IMG.H]	Byte, Float	Byte	[IMG.W+6 x IMG.H+6] + 10*4*(align(IMG.W+6),64) + 16 + [IMG.W x IMG.H]	37.7
7	Feature Descriptor	[265 x 4], [256 x 2]	[32 x 8]	Int, Byte	Byte	256 * 16 + KP*256*2+ KP*32	25
8	Orientation	[31 x 31]	1	Byte	Float	KP*359 + KP*4	21

Table legend:
IMG.W: image width
IMG.H: image height
CORN: # corners (input)
KP: # keypoints (input)

CONTACT

Fabrizio Petrini
Principal Engineer
fabrizio.petrini@intel.com

Josh Fryman
Senior Principal Engineer
joshua.b.fryman@intel.com

www.darpa.mil

Copyright © 2019 Intel Corporation. All rights reserved. Intel, the Intel logo, Xeon, and others are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others. This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government

Distribution Statement A – Approved for Public Release, Distribution Unlimited

THE ELECTRONICS RESURGENCE INITIATIVE