

計算科学研究センター
スーパーコンピュータ
COMA: PACS-IX の概要

COMA (PACS-IX)

- 筑波大学計算科学研究センターが平成26年4月より運用を開始するスーパーコンピュータシステム
(HA-PACSと平行して運用)
- T2K-Tsukubaの運転終了後に合わせて
- システム構成
 - 計算ノード: 汎用CPU + メニーコアプロセッサ (MIC)
 - ノード構成
 - CPU x 2: Intel Xeon E5-2670v2
 - MIC x 2: Intel Xeon Phi 7110P
 - Memory: CPU=64GB MIC=16GB
 - Network: IB FDR Full-bisection b/w Fat Tree
 - ノード数: 393
 - ピーク性能: CPU=157.2 TFlops MIC=843.8 TFlops
TOTAL: 1001 TFlops = **1.001 PFLOPS**
- システムベンダー: Cray Inc.

What is COMA ?

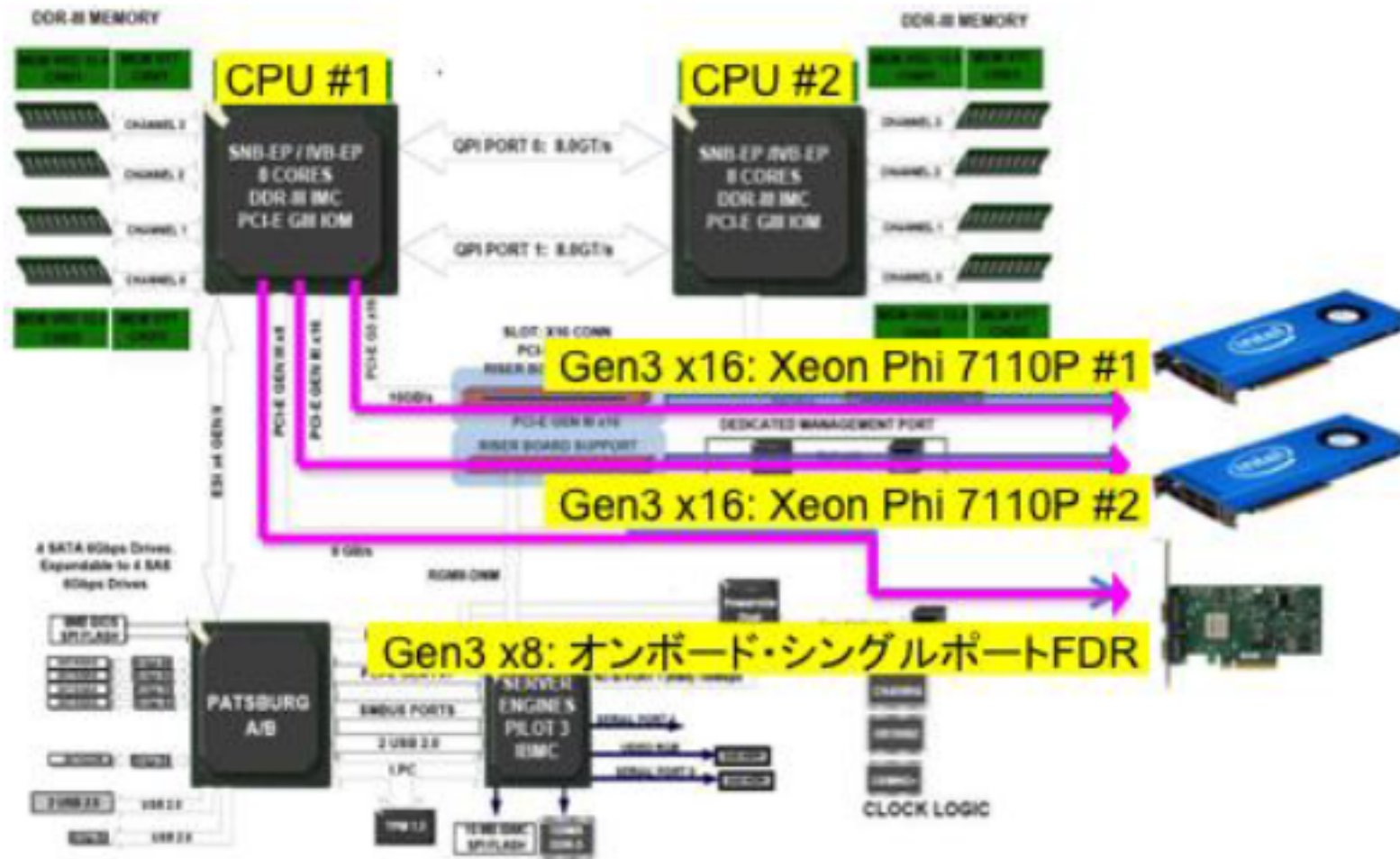
- Cluster of Many-core Architecture processor
- COMA＝「かみのけ座」
 - － 代表的な銀河団の一つ
 - － 銀河＝星の集まり(＝Many Core)
 - － 銀河団＝銀河の集まり(＝Cluster)
- 同時にPACSシリーズ第9世代のマシンとなるため、
"PACS-IX"のコード名を併用

COMAの計算ノード

- CPU (2基): Intel Xeon E5-2670v2 (Ivy Bridge)
 - 10 core/CPU, 2.5GHz
 - 200GFLOSP x 2 = 400GFLOPS
 - memory: 64GB
DDR3 1866MHz x 4chan x 2CPU = 119.4 GB/s
- MIC (2基): Intel Xeon Phi 7110P
 - 61 core/MIC, 1.1GHz
 - 1.0736 TFLOPS x 2 = 2.1472 TFLOPS
 - memory: 8GB x 2 = 16GB
GDR5 352GB/s x 2 = 704 GB/s
- Interconnection: InfiniBand FDR (on-board)
 - Mellanox Connect-X3
- Local HDD: 1TB x 2 (RAID-1)



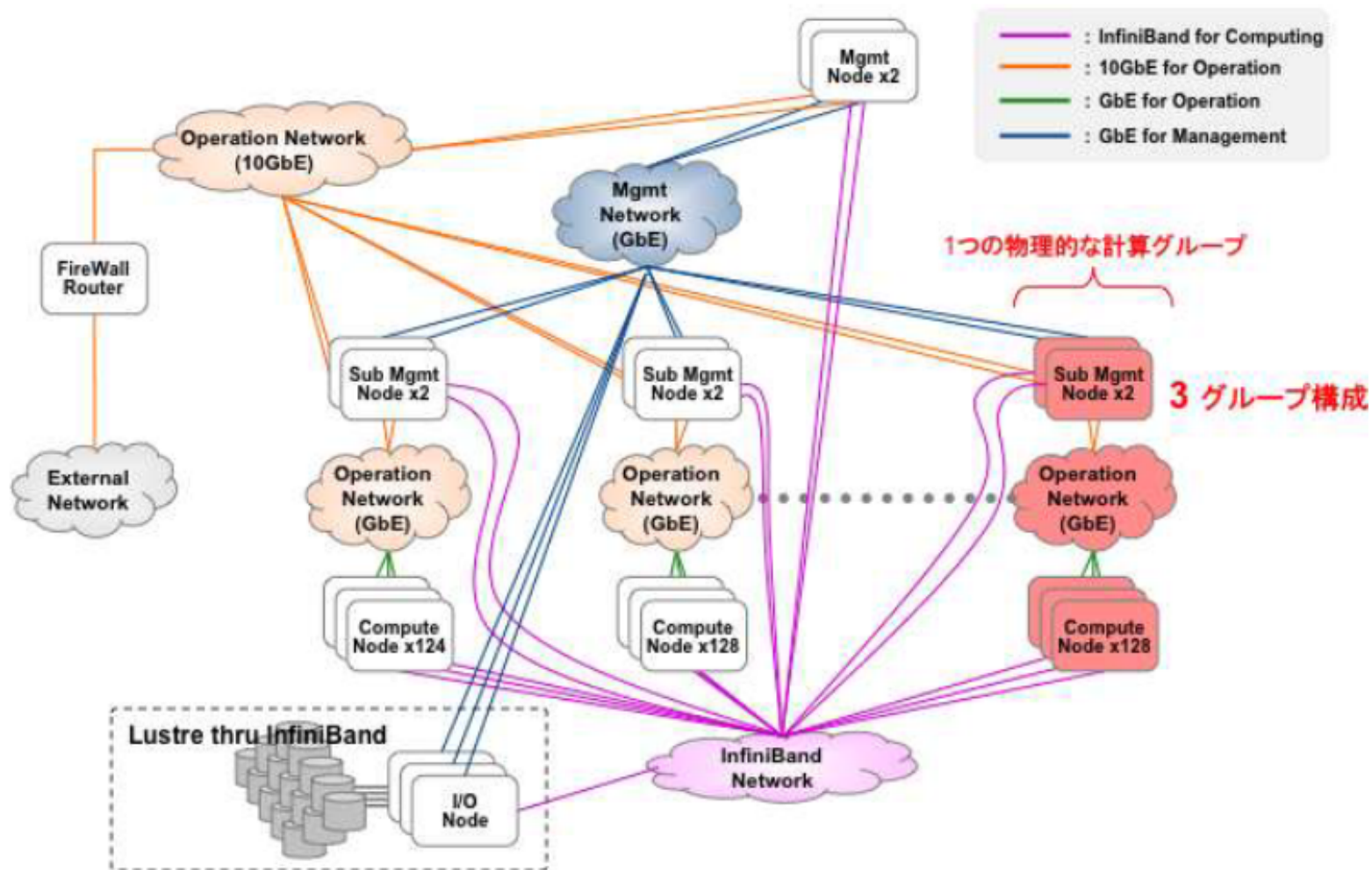
COMAの計算ノード



COMAシステム全体

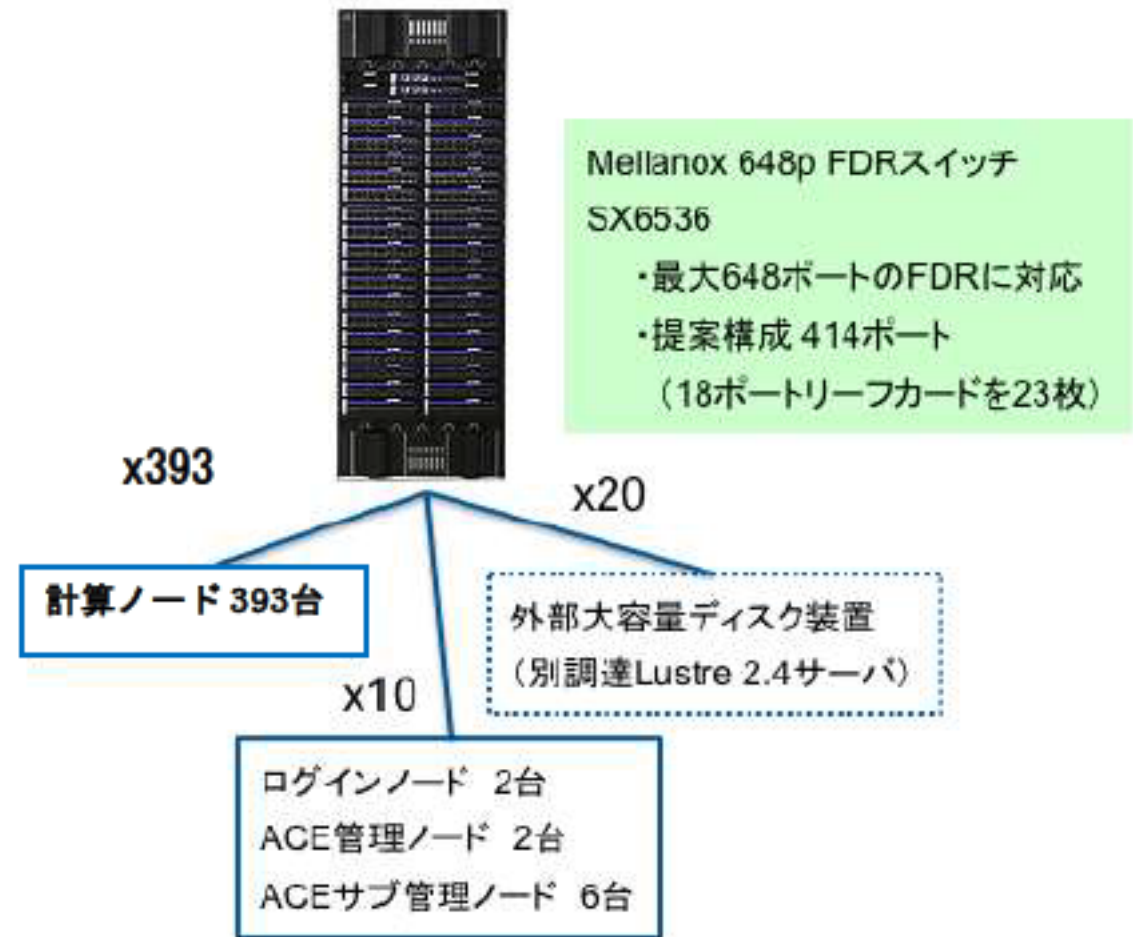
- ノード数: 393
 - ピーク性能
 - CPU: 157.2 TFLOPS
 - MIC: 843.8 TFLOPS
 - TOTAL: 1001 TFLOPS = 1.001 PFLOPS
 - メモリ容量
 - CPU: 25.1 TB
 - MIC: 6.3 TB
- Interconnection: Fat-Tree full-bisection b/w
 - Bisection bandwidth: 2.75 TB/s

COMAの全体ネットワーク構成

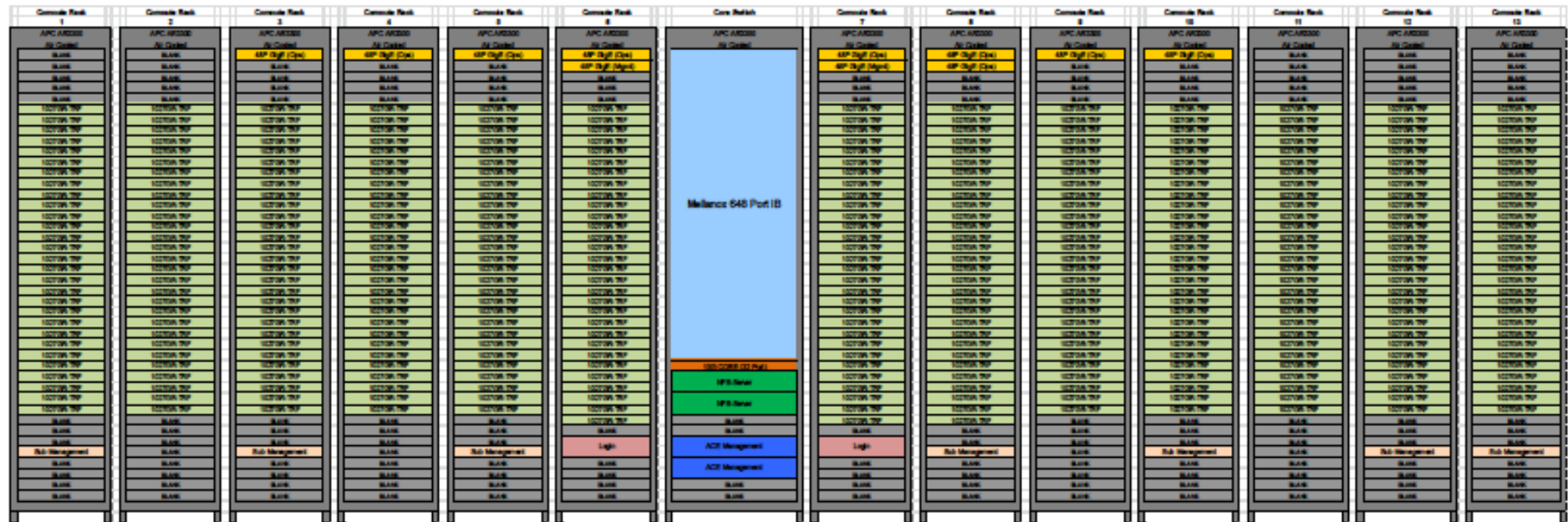


Interconnection Network

- Mellanox 648p FDR switch
 - 2.75TB/s bisection b/w
- Brocade ICX6650-32
 - 32port 10GigE SFP+



System Image



Storage

- Shared file server
 - RAID6 + Luster
 - OST: DDN SS8460 x 10
 - 4TB x 550基
 - OSS x 8
 - Controller: DDN SFA12K-40
 - InfiniBand FDR x8 (OSS)
 - User space: 1.5PB
 - Flat access from all computation nodes
- NFS server
 - 12TB for /home, etc.



ソフトウェア (OS, programming)

- OS: Red Hat Enterprise Linux (login server)
- OS: CentOS (compute node)
- Cluster management: ACE
- Job scheduler: SLURM
- Programming environment:
 - Intel Cluster Studio XE2013 (Composer/XE)
 - Fortran95/C/C++
 - Intel MPI

MIC (Many Integrated Core) の特徴

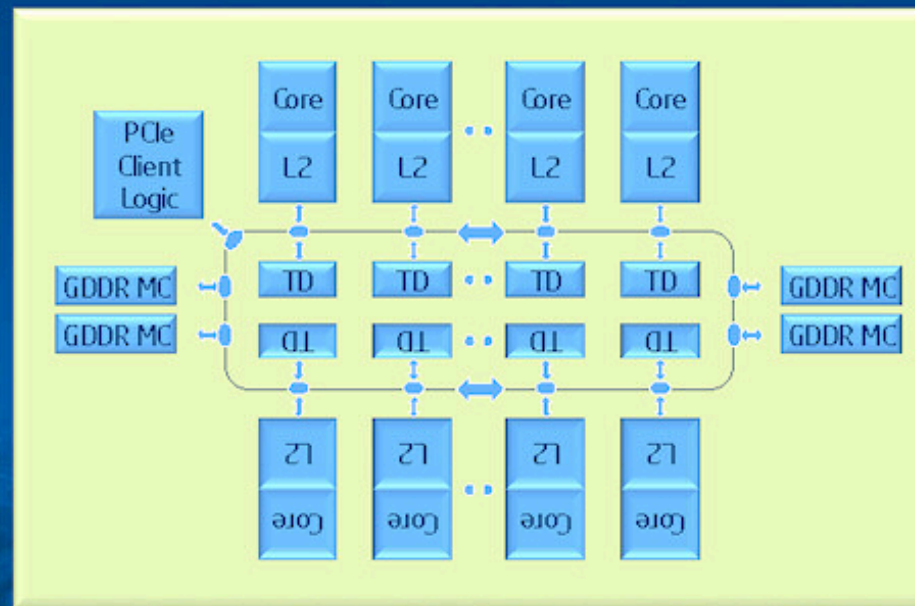
- Intelがmulti-coreプロセッサ（ハイエンド用としてはXeon family）と平行して近年開発している、多数のコアを1チップに搭載したプロセッサ
 - 一般的なmulti-core processorが数個～十数個のコアを搭載しているのに対し、MICは数十個（現行では約60個）のコアを搭載している
 - これまで試験的システムとしての第一世代チップ（コード名：KNF - Knights Ferry）の後、アクセラレータ型として実用的な第二世代チップ（コード名：KNC - Knights Corner）が開発され、現在このチップが“Xeon Phi” の名前で商品化されている
- 「コア数が数倍あるのだから、性能も比例して上がるだろう」という期待は必ずしも正しくない

MICの特徴

- 商用化KNC(例:Xeon Phi 5110P)の特徴
 - 単体のCPUとして stand alone で動作するのではなく、GPUのように PCI Express バスでCPUに接続されるアクセラレータである
 - 1カードに1台のXeon Phiチップとメモリを搭載したPCI Expressカードとして提供(GPUに似ている)
- MICのコアの特徴
 - 先進的なXeonのコアとは大きく異なり、基本的に非常に単純な構造・分岐予測システム等を持ち、2～3世代前のIntel CPUに近い(Pentiumレベル)。
 - このため、整数演算性能や分岐の多い一般プログラムの実行は非常に遅い(Linux等が走るが、やはり遅い)。
 - 一方、浮動小数点演算性能のピーク値は非常に高い。512bit SIMD命令が用意されている (AVX-512)。

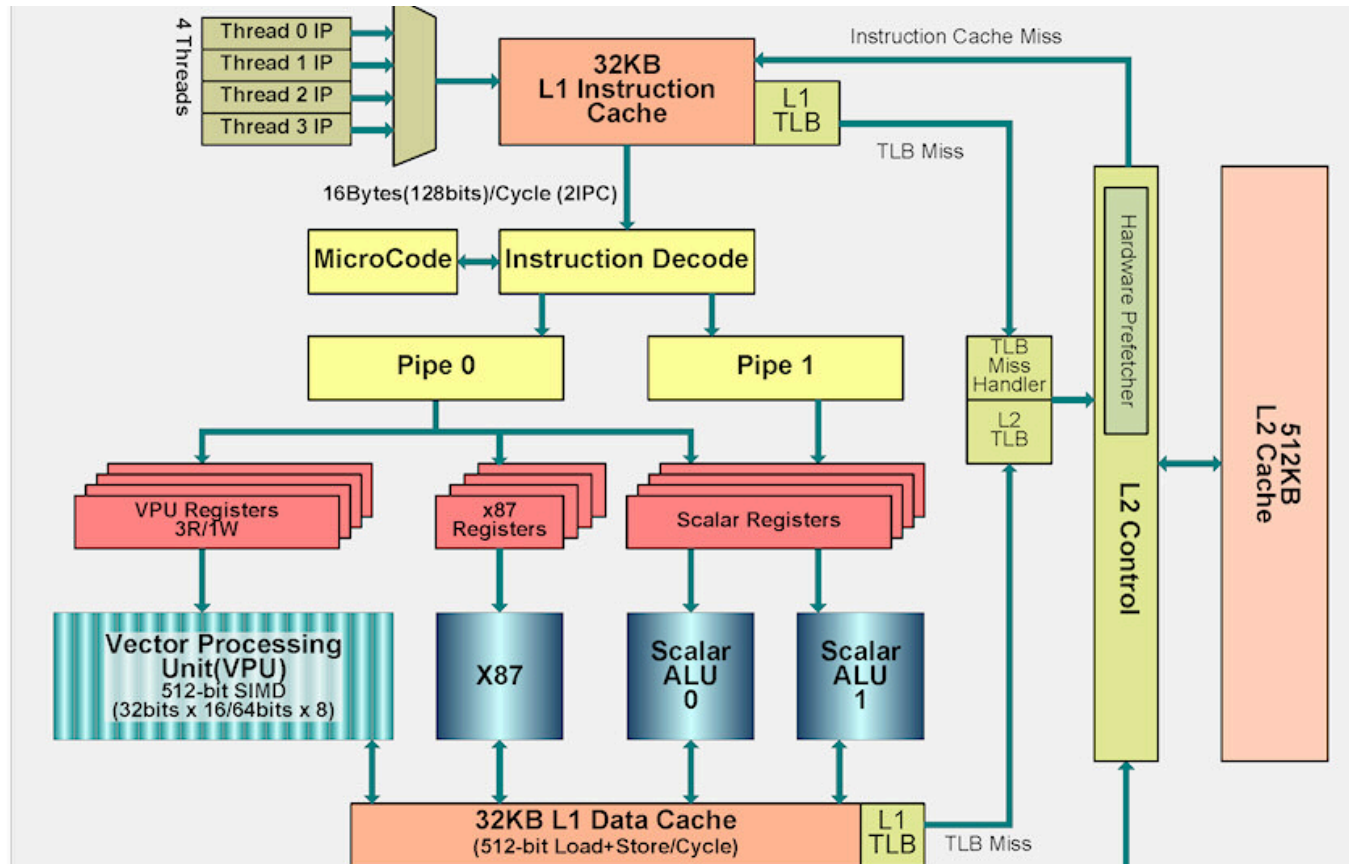
MICアーキテクチャ

Knights Corner Micro-architecture



- Ring Busによって60個のコアを接続
- メモリはGPUと同じGDRメモリ、複数チャンネルでバンド幅を増強
- L1+L2キャッシュを各コアに配置、L1キャッシュは占有だがL2キャッシュはコア間でコヒーレンス制御される
- TD: Tag Directory (cache制御用)

各コアのアーキテクチャ



- 各コアは4つのハードウェアスレッドを持つことができ、hyper-threadingで動作
- 演算器はコア当たり1セットだが、レジスタやプログラム・カウンタはスレッド毎にある
- 512bit SIMD (16-way for SP, 8-way for DP), FMA (AVX-512)

MICのプログラミング

- MIC(KNC)はシステム全体の中ではアクセラレータの位置づけだが、x86命令セットなのでLinuxが動作する
- 3つのプログラミングモデル
 - Native Mode: MIC上のLinux上で直接プログラムを実行
 - OpenMP等でスレッド並列化されたアプリが動く
 - ハードウェアスレッドを使うと最大240スレッド
 - I/O装置(HDD等)を持っていないので、ホストノードのHDD等をNFSマウントしてアクセス
 - MPIも動き、ホスト上のMPIプロセスとの通信も可能
 - Offload Mode: ホスト上のCPUからプログラム上でoffload指定された部分をMIC上で実行
 - IntelのMIC用拡張コンパイラ (Composer XE) を使う
 - デバイス(MIC)で実行される部分を明示
 - Offload部にOpenMP等を書けばMIC上の並列実行を記述可能
 - Symmetric Mode: ホスト上のCPUコアとMIC上のコアをフラットに扱いMPIプロセスを割り当てる

COMAのオペレーション

- 3つのパーティション(ジョブキュー)を用意
 - CPU partition (HPCI、学際共同利用)
 - 各ノードの20 coreのXeonのうち16coreを使用
 - これまでのmulti-core用プログラム(+MPI)を実行
 - MIC partition (学際共同利用)
 - 各ノードのXeonの 4core + 2MICを使用
 - Offload model が中心
 - Mixed partition (大規模一般利用、学際共同利用)
 - ノード単位でCPU+MICを丸ごと利用可能
 - ハイブリッドプログラム、MICのnative mode実行、Symmetric Mode

COMAの運用

- 学際共同利用
 - CCSが継続している先進的計算科学／計算工学研究支援プログラム
 - 無償利用(要申請＋審査、H26年度申請は締め切り)
 - CPU, MIC, Mixed
- HPCI
 - 文科省が全国主要スパコンをsingle sign onで利用可能としている共同利用プログラム
 - 無償利用(H26年度申請は締め切り)
 - CPU
- 大規模一般利用(有償)
 - ノード単位で有償利用
 - 公募は間もなく
 - Mixed
- マシン利用開始は4/15の予定