今後の高性能計算システムについて
～アクセラレータの立場から

理化学研究所　基幹研究所
システム計算生物学研究グループ

泰地　真弘人
taiji@riken.jp

今後のHPC向けプロセッサ開発

- HPC向け（特にトップエンド）とコモディティの間のギャップは拡大
- HPC向けに必要な要素
  - 高並列プロセッサ・アクセラレータ
  - チップ積層・貫通配線
  - ネットワーク統合
- プロセッサよりネットワークの比重が高くなる。

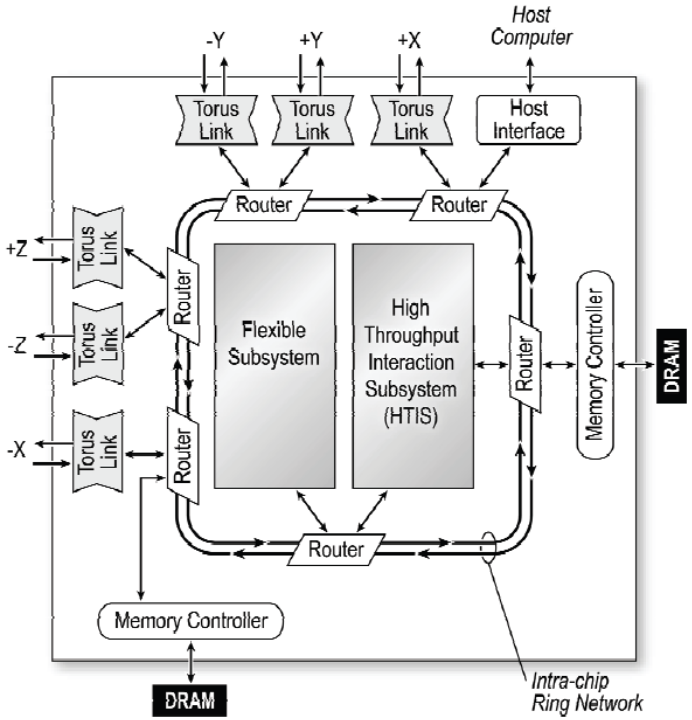
  HPC専用が必要な理由も、ネットワークになるのでは
- 次世代スーパーコンピュータの利点も、既にその側面が大きい。但し、HPC専用化の利点を十分に活かし切ってはいない。

ネットワークの課題

- 例：分子動力学計算
  Strong Scalingが要求される
- ざっくりとした計算量～50,000演算/粒子
- 10^5粒子の系
- 5 GFLOP/step
- 実効5TFLOPSの計算機で、
  1msec/step = 170nsec/day
  実現可能
- 実効5PFLOPSの計算機では
  1μsec/step = 200μsec/day???
  難しい、けどこれを実現したい

Anton

- D. E. Shaw Research
- 専用パイプライン
  + 汎用コア
  + 専用ネットワーク
- プロセッサも重要だが、通信の最適化でどこまでボトルネックを解消できるかを示したことが重要



	GROMACS time		Anton time	
	small cutoff (9Å) large mesh (64^3)	large cutoff (13Å) small mesh (32^3)	small cutoff (9Å) large mesh (64^3)	large cutoff (13Å) small mesh (32^3)
Nonbonded forces				
Range-limited forces	111 ms (61%)	308 ms (88%)	1.8 μs (3%)	3 μs (13%)
FFT & inverse FFT	29 ms (16%)	3 ms (1%)	38 μs (66%)	12 μs (50%)
Mesh interpolation	19 ms (10%)	18 ms (5%)	10 μs (17%)	5.5 μs (23%)
Correction forces	7 ms (4%)	6 ms (2%)	2 μs (3%)	2.5 μs (10%)
Bonded forces	9 ms (5%)	9 ms (2%)	5 μs (9%)	5 μs (21%)
Integration	7 ms (4%)	7 ms (2%)	3 μs (5%)	2.5 μs (10%)
Total	181 ms (100%)	351 ms (100%)	58 μs (100%)	24 μs (100%)

R. O. Dror et al., Proc. Supercomputing 2009, in USB memory.

今後のネットワークに望まれること

- μsecオーダーのタイムステップまでの強スケーリング
- 最後はネットワークが性能の限界を決める
- ネットワークのお約束ごとが煩わしい

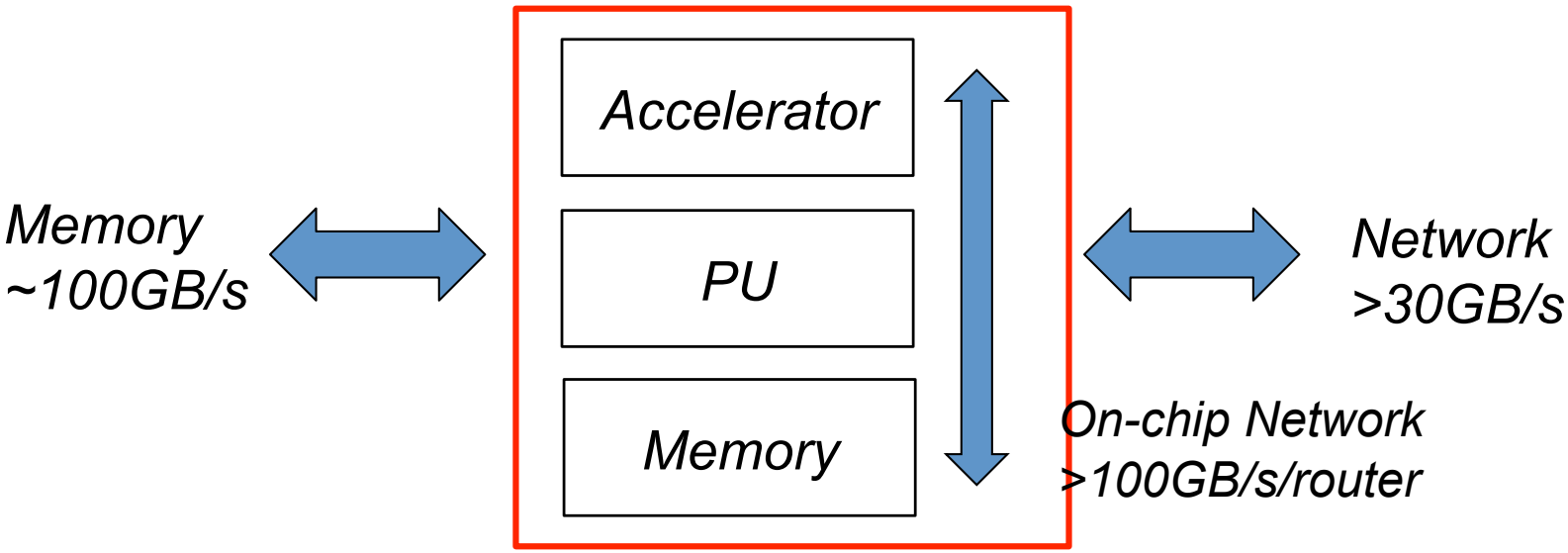
  HPC向けなら、ネットワークは1～数プロセスが占有すると思ってもかまわないはず
- 主記憶がうっとうしい

  例えば、MDはオンキャッシュで十分。一般的にも、通信データはキャッシュ・レジスタから出て、キャッシュで受け取ってというのが最も効率がよい場合が多いはず。
- 短いパケットで低レイテンシで通信したい

Back to PAX?

Future Directions (1)

- 専用機も汎用機も、ネットワークの統合は必須
- Platform for Accelerators
  - General-purpose processor cores
  - Cache or local memory
  - Modest memory access
  - Fast, low-latency on-chip and off-chip networks

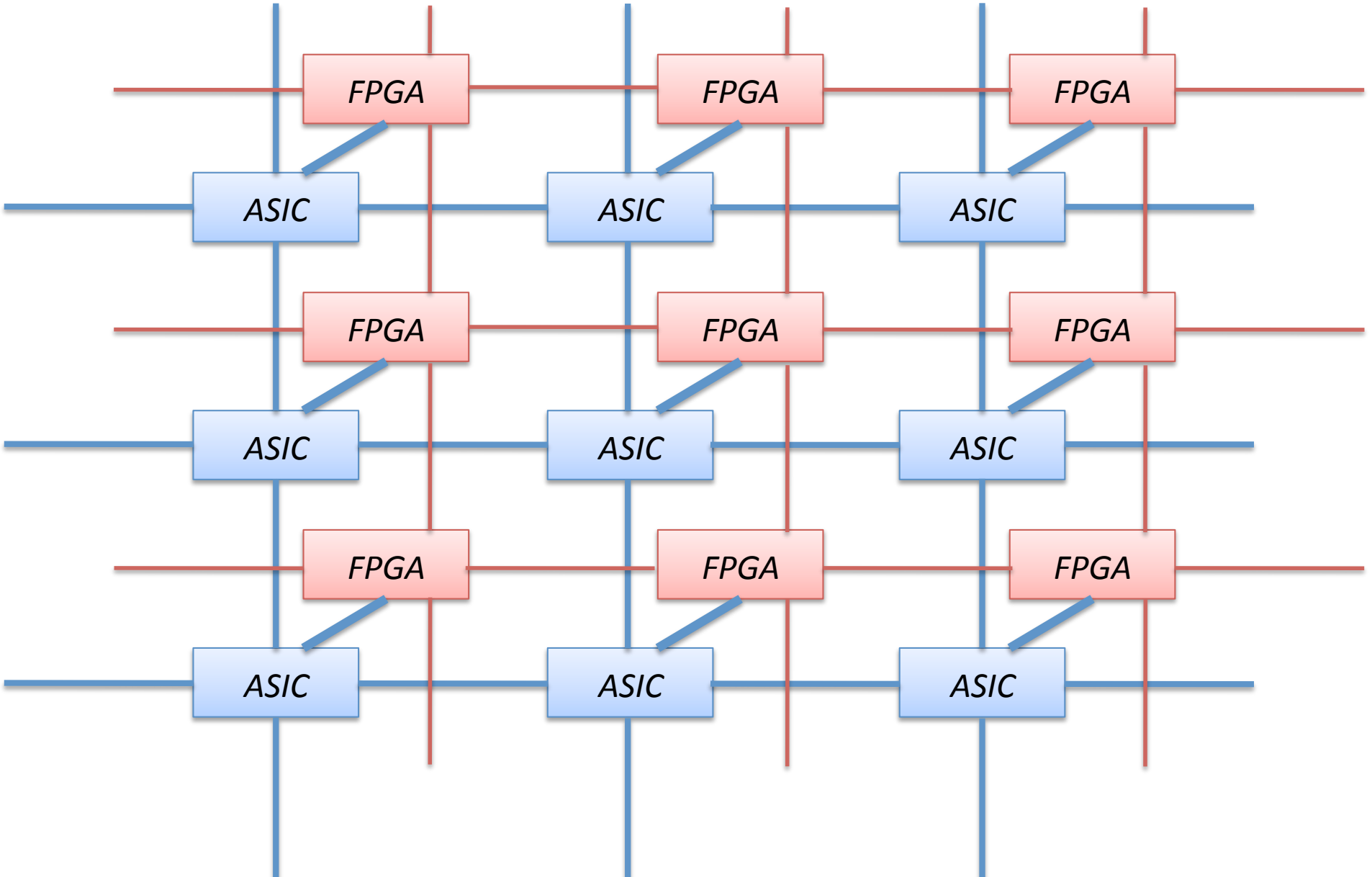


MDGRAPE-4

- 今年から開発予定
- 元々はもう少し汎用度を上げたタイルプロセッサを開発するつもりで準備してきたが…ネットワーク周りを強化する必要があり、時間がない！

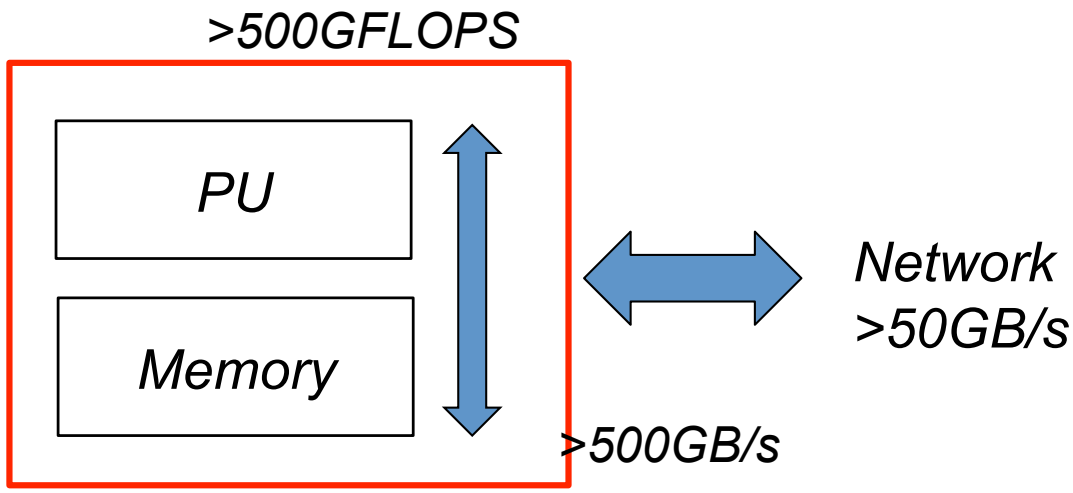
- MDGRAPE-3が素人臭かった点を直す
- MD専用でAntonに対抗
- ネットワーク帯域を上げる
- 短距離力向け：1方向＞5(+5)GB/s, 6方向
- 長距離力向け：Multigrid/FMM向け専用ネットワーク？
- 汎用プロセッサの取り込み
- キャッシュメモリ等の一部採用

MDGRAPE-4 System



Future Directions (2)

- High Memory Bandwidth System
  - シングルチップ・または貫通配線による「1チップBG/L」
  - B/F～1
  - B/F～0.1 for remote node



専用アクセラレータか、汎用か？

- 専用のメリット
  - ちゃんと作れば、「究極の性能」
  - 少ないレジスタ・メモリアクセスで、多くの演算
  - SIMDプロセッサの場合に重くなる、Reductionの問題を軽減できる
  - 演算精度・転送の効率化による、消費電力減

- 専用のデメリット
  - 高い開発費
  - 今後は、汎用部分を取り込む必要がある
    → 開発の難度が上がる

- GPU
  - まじめにHPCをやるなら、汎用バス経由の接続に頼っていてはだめなのでは？見かけのコストは下がりますが。