超高メモリバンド幅マルチメディアプロセッサアーキテクチャを目指して

~3次元積層技術によりアーキテクチャ設計空間のさらなる拡大を~

小林広明

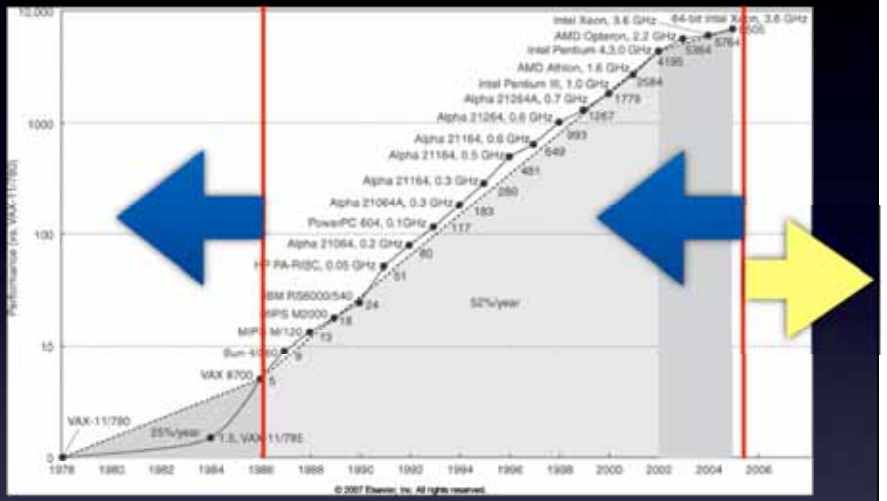
koba@isc.tohoku.ac.jp
東北大学

(小柳チーム)

2010年10月1日

背景：プロセッサアーキテクチャ設計のトレンド

- 大昔(~1985)
  - 言語とハードウェアのセマンティックギャップを埋めよう!
  - CISC時代
- これまで。。。(~2005)
  - ★ シンプルに行こう! /
  - ★ RISC/Superscalar時代
- 現在＆これから!? (2006~)
  - ★ アプリ・言語レベルの並列性とハードウェアの間のギャップを埋めよう!
  - ★ Wide-SIMD/ベクトル&マルチ・メニーコアの時代!?
    - ✓ Vector (Wide-SIMD) ISA
    - ✓ Vector FUs/Register Files
    - ✓ Large on-chip memory
    - ✓ High Memory Bandwidth
    - ✓ Multiple Vector/SIMD cores

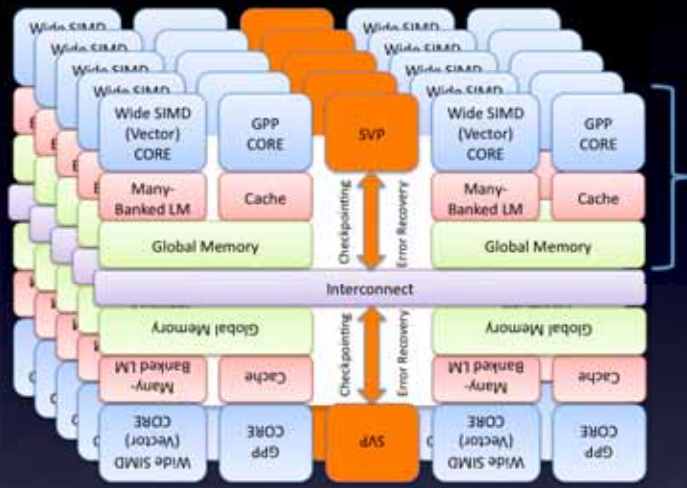


Hennessy&Patterson CAQA 4thEd., 2006

メディアアプリはベクトル処理の宝庫!?

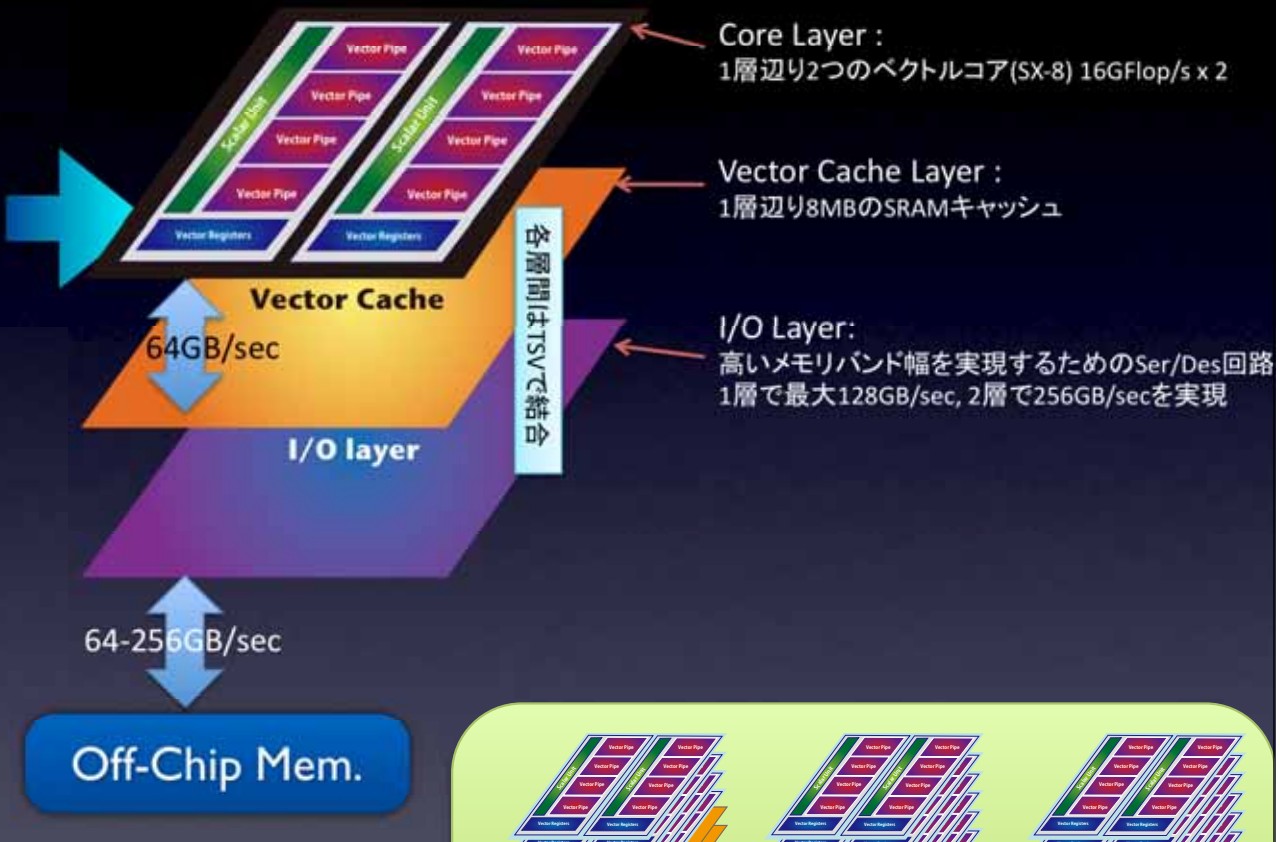
Benchmarks	Application Domain	Vecterization Ratio	Vector Length	Benchmark Suite
Sphinx3	Speech recognition	99.42%	4096	Parsec
FaceRec	Face recognition	98.17%	173	Parsec
Raytrace	Animation	99.64%	1080	ALPbench
Vips	Image processing	98.35%	79	ALPbench
M x M	Matrix Mult.	98.90%	1000	
V x M	Vector Mult	99.10%	1000	

3次元積層技術によりアーキテクチャ設計空間を広げよう



- ✓ Vector (Wide-SIMD) ISA
- ✓ Vector FUs/Register Files
- ✓ Large on-chip memory
- ✓ High Memory Bandwidth
- ✓ Multi-vector/SIMD cores

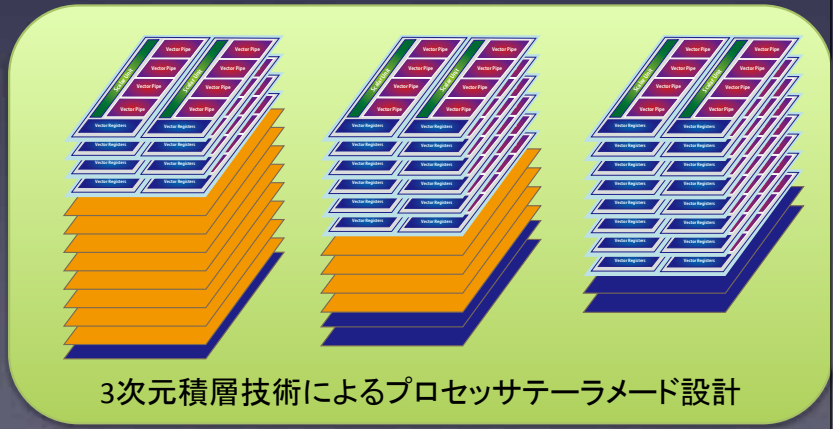
2次元設計の限界



Core Layer :
1層辺り2つのベクトルコア(SX-8) 16GFlop/s x 2

Vector Cache Layer :
1層辺り8MBのSRAMキャッシュ

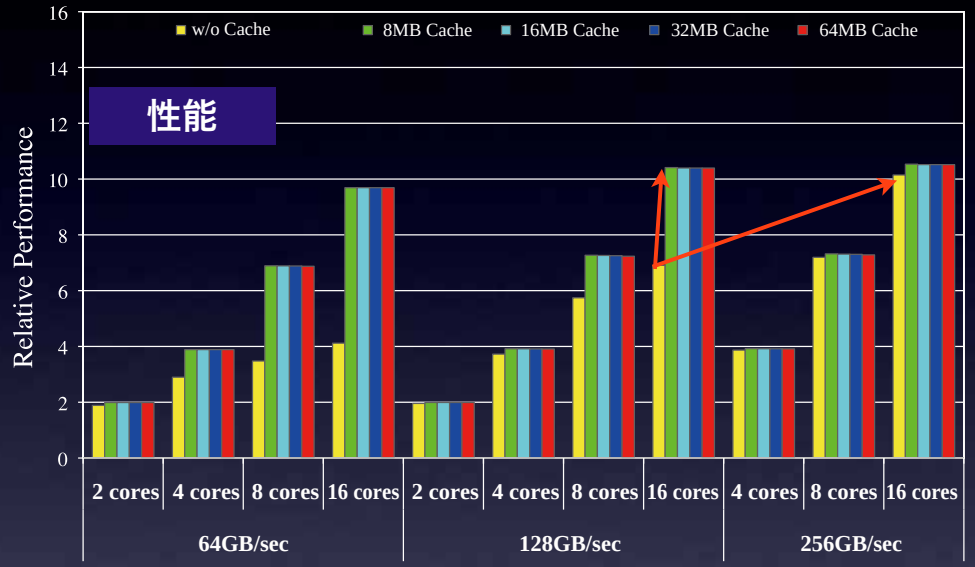
I/O Layer:
高いメモリバンド幅を実現するためのSer/Des回路
1層で最大128GB/sec, 2層で256GB/secを実現

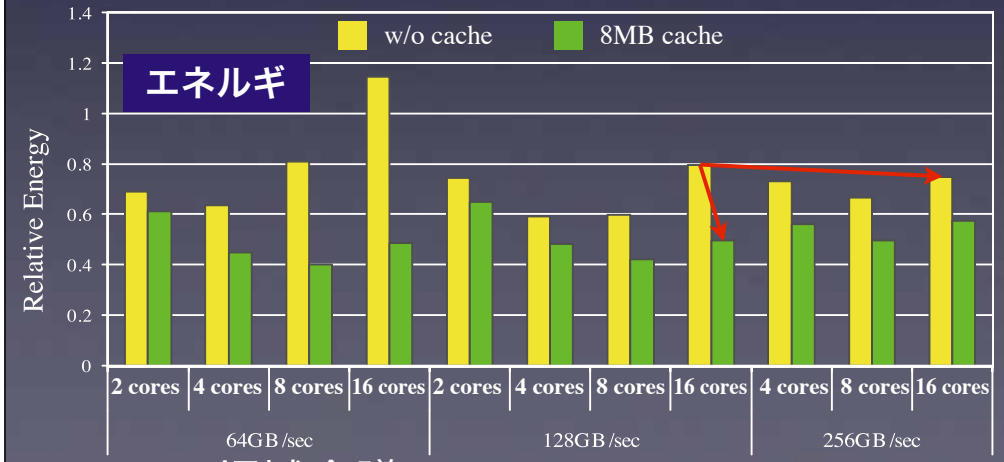


3次元積層技術によるプロセッサテーラメード設計

アプリ毎に適切なトレードオフ設計が可能

Energy-Conscious Architecture Configuration





Off-Chip BW vs. On-Chip Cache?

キャッシュによる性能向上はオフチップ
メモリバンド幅による性能向上に等しい

16Cores + 256GB/sec

∧ 性能

16Cores + 128GB/sec+8MBCache

キャッシュによる性能向上は電力効率が高い

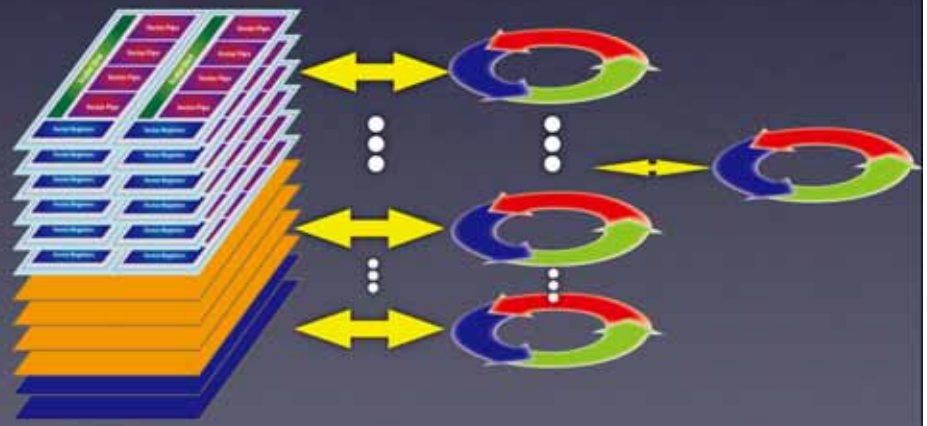
16Cores + 256GB/sec

∨ エネルギ

16Cores + 128GB/sec+8MBCache

ディペンダブルアーキテクチャ設計に向けて

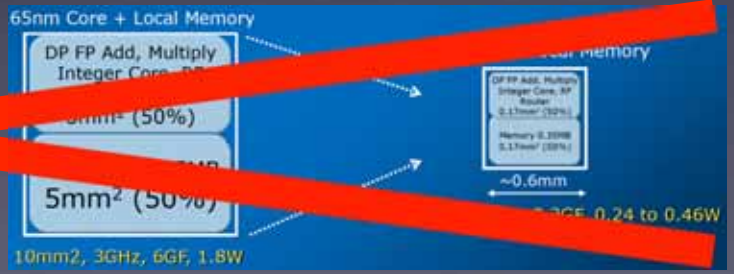
- 目標：QoS（性能、RAS）の維持
- Self-Aware System: adaptive & self-healing architecture の実現
  - power/resource proportional computing
  - Goal-oriented computing
  - Gracefully reducing computing
- Basic Cycle for Adaptive & Self-healing Functionality
  - 環境監視（パフォーマンス/リソース/ヘルスモニタリング)
  - 実行環境評価
  - 適応動作
    - ハードウェア再構成
    - OS/runtime support連携
- Control Granularity（3次元レイアレベル）
  - Processor Core Level
  - On-Chip memory level
  - I/O (Memory BW control) Level



まとめ

- ★ 安定したプロセス技術＋新3次元実装技術で、 アーキテクチャ設計に新時代を！
  - Only one & Number oneの技術を目指して
  - テクノロジ競争(Mooreの呪縛?!)からの脱却
- ★ 時代はベクトル/Wide-SIMD＆マルチコアコンピューティングへ。
  - 情報爆発時代に突入し、どんどん広がる高性能マルチメディアプロセッシングの適用範囲
    - マルチメディア処理、マルチメディア合成(CG,AR ...)、マルチメディア理解（データマイニング） ・・・
- ✓ Computation-Intensive/Data-Level Parallelism
- ✓ Highly-Efficient Vector & Parallel Computing
- ✓ Energy-proportional/Adaptive/Self-Healing Computing

(High Volume)	45nm	32nm (2010)	22nm (2012)	16nm (2014)	11nm (2016)	8nm (2018)	5nm (2020)
Transistor density	1.75				1.75	1.75	1.75
Frequency scaling	15%	10%	8%	5%			
Vdd scaling	-10%	-7.5%			-1.5%	-1%	-0.5%
Dimension & Capacit		0.75	0.75	0.75	0.75	0.75	0.75
scaling/micron	1X Optimistic to 1.43X Pessimistic						



Source: Intel