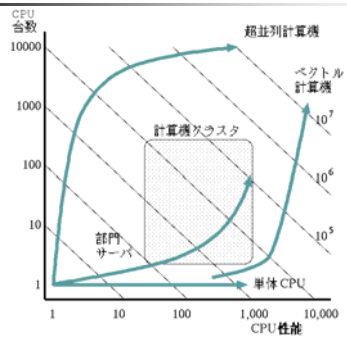
平成19年度(11/21版)

並列分散システム論 Part II

福井大学大学院工学研究科　情報・メディア工学専攻
計算機・通信講座　計算機アーキテクチャ分野
森眞一郎

並列・分散システムの実現
- 計算機クラスタ
- 分散共有メモリシステム
- Grid　(??)

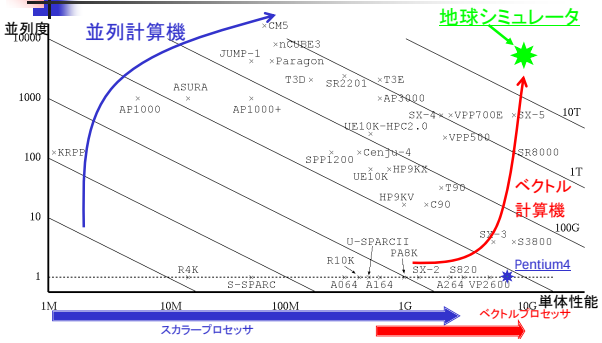
計算機クラスタの位置付け



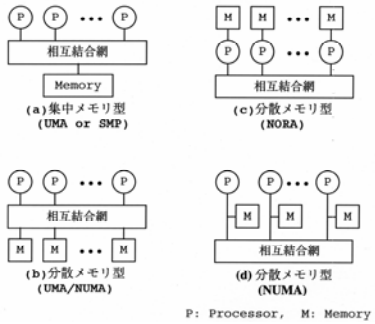
「計算機クラスタ」の追い風は？

- CPUの高速化/低価格化…… Alpha/Pentium (PCの追い上げ)
  - 8GFlopsの壁???……Alpha(612MHz)の6倍，UltraSPARC(336MHz)の12倍
    9.6GF??　　Pentium4(3.06GHz)の1.5倍
    ⇒ NEC SX-6 (8GF,500MHz) , SX-7 (11.41GF), SX-8 (16GF,2GHz)
- 通信媒体の高速化
  - 2Gpbsの壁??? ….多重化（波長,位相,振幅），並列化，光のWave Pipeline
    ⇒ 10Gbps メタル
- 標準通信ライブラリの普及
  - Posix Thread, OpenMP, HPF, MPI, AM, FM, PM, …..
- 米国（だけ??）での並列計算機メーカの不振…….(最近は??....CMPで挽回？)
  - 地球シミュレータ…… 5120プロセッサ，40TF (2002.03@海洋科学技術センタ 横浜)
  - Cray Inc. …… 単体性能12.8GFlopsのベクトルプロセッサを発表(2002.11)
    - 3.2GFlops のスカラプロセッサ4台＋ベクトル処理向けに強化した同期機構
      ⇒仮想ベクトル処理
    - 4096プロセッサで52.4TFlops（未稼働)、　2010年までに 1 PetaFlops を目指す
  - Blue Planet …. 2005年を目処にIBM Power5(10GF) x 16,384 =150TFlops

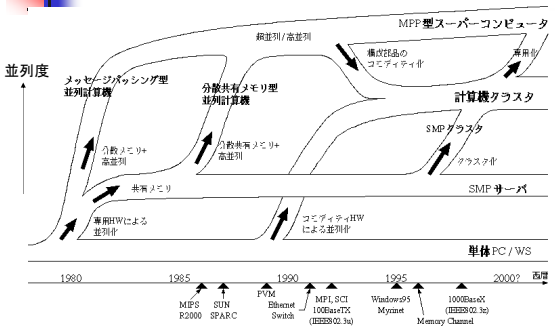
いつまで続くCPUの高速化



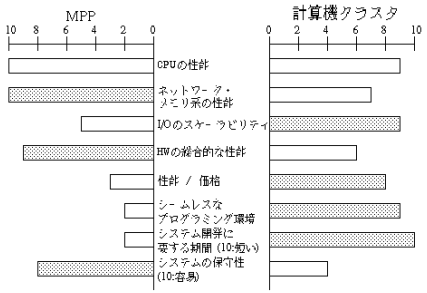
並列計算機の分類



計算機クラスタへの道程



MPPと計算機クラスタの比較



TOP500 Supercomputer @2007.11

- 世界一位はBlueGene/L @US DOE/NNSL/LLNL
  - 212,992 プロセッサ、478.2TF (理論最大596.47TF)
    - IBM PowerPC 440 700 MHz (2.8 GFlops)
- Top500の内
  - 80% が　Cluster
  - 54% が　GigabitEthernet, 24%が　Infiniband
  - 約1/5が Intel Clevertown Quad Core, 残りの大半がDual core
- 日本の位置づけ
  - 日本1位(世界16位)　Tsubameクラスタ@東工大
    - 11664proc, 56,430TF (102,021TF)
  - 日本2位(世界30位)　地球シミュレータ
    - 5120proc(SX-6 8GF), 35,860TF (40,960TF)

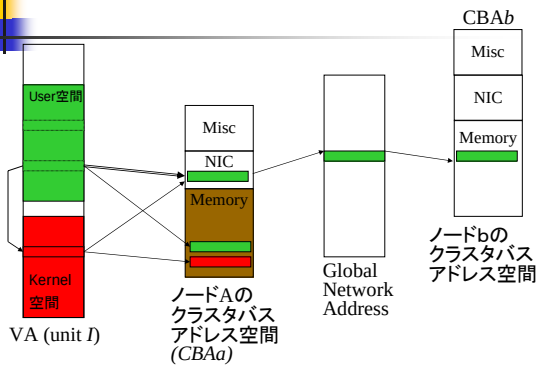
ネットワークの急速な高速化

- 通信媒体自体の高速化
- 通信ソフトウェアの高速化
  - 0コピー通信
  - remote read/write (put/get)
  - 軽量化プロトコル

通信媒体自体の高速化

- 汎用の通信媒体として
  - 10Base2/5⇒10BaseT⇒100BaseT⇒GbE⇒10GbE
    - バス型結線 スター型結線 …… 相互干渉の軽減, 高周波設計の容易さ
    - 電気 光 …… 物理的な距離に関する制約の緩和
  - USB(Universal Serial Bus), IEEE1394バス
    - 家庭内ネットワークコンピューティングへの足掛かりになるか??
    - 480Mbps@USB2.0 ⇒ 4.8Gbps@USB3.0 (IDF2007Fall)
- 専用の超高速通信媒体として
  - Myrinet(Myricom社) 通信用コプロセッサを搭載
  - Memory Channel(DEC) 仮想共有メモリ環境を提供する PCI バス用
  - SCI (IEEE1596) CC-NUMA環境
  - Synfinity NUMA(Fujitsu) CC-NUMA環境 + 1.6Gbps
  - QsNet (Quadrix) VA-to-VA Remote DMA, 340MB/s
  - Advanced Switching PCI-Expressベースの拡張バスシステム

Zero-copy通信



Myrinet

- USC/ISI+Caltecの共同開発(ATOMIC) の commercial version
  - 専用通信プロセッサ LANai
  - 高速パケット交換スイッチ (2Gbps + 2Gbps)x2@MyrinetXP
- Application Programming Interface や LANai chip 仕様 の公開
  - 多くの Hacker達が競って最適化！！
  - UC Berkeley(NOW), Illinois(Fast Messages), RWCP(PM)

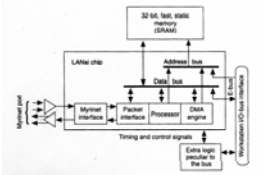


Fig6 通信ボードのブロック図 (文献 [1])

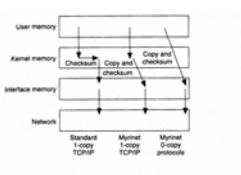
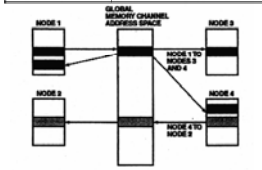


Fig7 Zero-copy プロトコル (文献 [1] より引用)

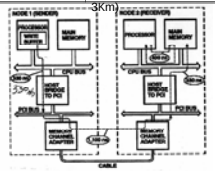
Memory Channel

- 仮想共有メモリ環境を提供する PCI バス用通信ボード
- 異なるノード上のプロセス間通信が, load/store で実現可能

世代	転送速度 (ユーザプロセス間sustained)	遅延 (store to load)	ハブの構成 (cabling)
第一(1996～)	77MB/s (66MB/s)	2.9μs	バス構造(電線4m)
第二(1997～)	97MB/s (88MB/s)	2.2μs	スイッチ(電線10m or 光 3Km)



クラスタワイドなアドレスマッピング



Latency Contribution

汎用ネットワークの追い上げ

超並列計算機のネットワーク性能

製品名	メーカ	発表年	Link速度
CM-5	T.M.	1992年	1.06Gbps
Paragon	Intel	1992年	1.6Gbps
T-3D	Cray	1993年	1.2Gbps
T-3E	SGI	1995年	4.8Gbps
SR2201	日立	1996年	2.4Gbps
VPP700E	富士通	1997年	4.8Gbps

製品名	メーカ	発表年	Link速度
SR8000	日立	1998年	1.0GB/s
SX-5	NEC	1998年	8(?)GB/s
Cenju-4	NEC	1998年	0.8GB/s
VPP800	富士通	1999年	1.6GB/s
AsamA	NEC	2002年	12.8GB/s
地球Sim	NEC	2002年	12.3GB/s

ギガビット級ネットワーク@2004

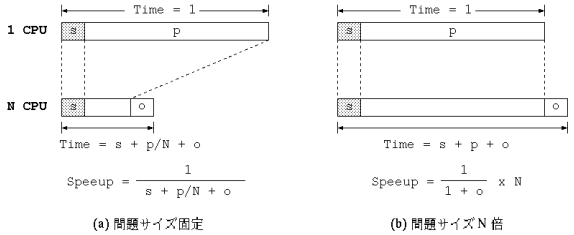
名称	規格/企業	単方向Link速度	通信の信頼性@物理層
Fibre Channel	ANSI X3 T11	2.12Gbps	Reliable
Memory Channel2	HP (DEC)	1.06Gbps	Reliable
Myrinet	Myricom	2.56Gbps	Reliable
Infiniband	Infiniband(Intel他)	2.5～30Gbps	Reliable
10Gigabit Ethernet	IEEE 802.3ae	10Gbps	Unreliable
(参考)VisA PRO Link	富田研	10Gbps	Reliable/Unreliable

総務省の資料2003/05/07 によると数年後には40GbE/100GbEも登場！！

プロセッサ間通信の遅延

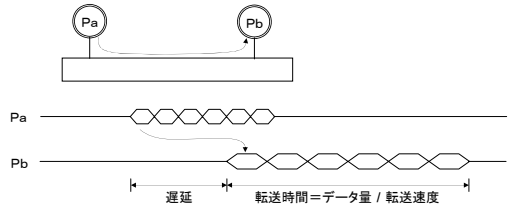
NIC	Latency	RATIO	I/F
DIMMnet-1	225ns(??)	1	PC133DIMM
ELAN	2us	10	66M64bPCI
Memory Channel	2.2us	10	33M32bPCI
SCI	2.3us	10	66M64bPCI
VIA on CLAN1000	3.5us	16	33M64bPCI
PM on Myrinet	7.5us	33	33M32bPCI
GM on Myrinet2000	7.6us	33	66M64bPCI
PM on GbE	24.1us	107	33M32bPCI
AsamAカスタムLSI	300ns	1	ItaniumII bus

スピードアップに関する2つの視点



並列処理オーバヘッド

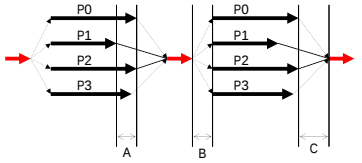
（1） 通信オーバヘッド



[技術的背景] 転送速度の向上は著しいが、遅延時間の短縮は極めて困難

並列処理オーバヘッド

（2） 負荷のアンバランスと同期オーバヘッド



A: 負荷のアンバランスによるオーバヘッド

B+C: 同期オーバヘッドの例

同期操作のオーバヘッド削減

- スピンロック
  - Test and Test&Set……同期待ちトラフィックの軽減

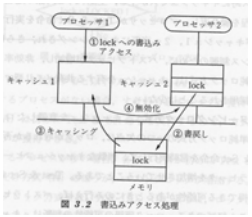


図 3.2 書込みアクセス処理

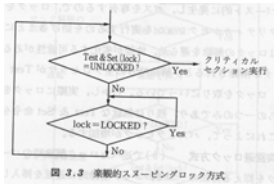


図 3.3 楽観的スヌーピングロック方式

集中型バリア同期機構の例

★カウンタベース

★再利用可能

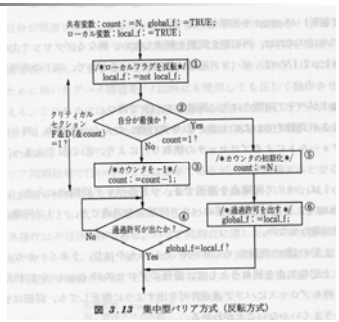


図 3.13 集中型バリア方式（反転方式）

分散バリア同期機構の例

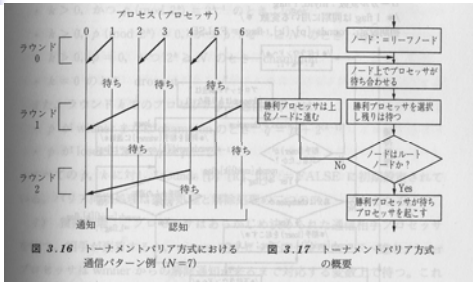


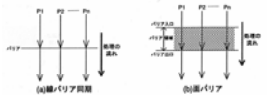
図 3.16 トーナメントバリア方式における通信パターン例（N＝7）
図 3.17 トーナメントバリア方式の概要

バリア同期モデルの分類

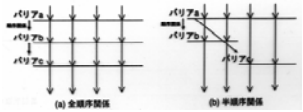
分類項目				バリア同期モデル
線／面	強制参加／任意参加	全順序関係／半順序関係	オーバラップ不可能／可能	
線	強制参加	全順序関係	不可能	(普通の)バリア
	任意参加	全順序関係		SBM
		半順序関係		HBM, LDBM, FDBM
面	強制参加	全順序関係	不可能	Fuzzyバリア
			可能	重複可能バリア
	任意参加	全順序関係	不可能	文献[104]で提案
			可能	Elasticバリア
		半順序関係	不可能	文献[104]で提案
			可能	文献[104]で提案

バリア同期モデルの分類

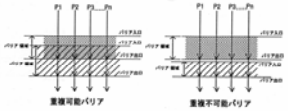
1）線バリア vs. 面バリア



2）全順序 vs. 半順序



3）重複可能 vs. 重複不可能



4）強制参加 vs. 任意参加

Transactional Memory --Retry ベースの投機的な同期処理--

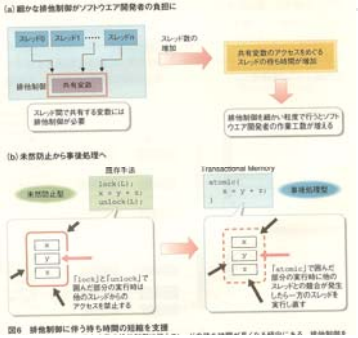


図6 排他制御に伴う待ち時間の短縮を支援

「マルチコアが変えるソフトウェア開発」日経エレクトロニクス 2006.10.09, No. 936, p.67--87 より引用

Transactional Memory --Retry ベースの投機的な同期処理--

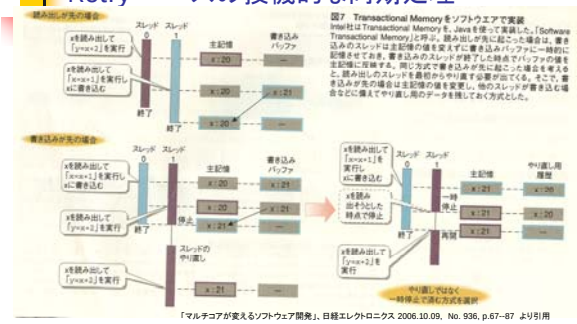


図7 Transactional Memoryをソフトウエアで実装 Intel社はTransactional Memoryを，Javaを使って実装した．「Software Transactional Memory」と呼ぶ．読み出しが先に起こった場合は，書き込みのスレッドは主記憶の値を変えずに書き込みバッファに一時的に記憶させておき，書き込みのスレッドが終了した時点でバッファの値を主記憶に反映する．同じ方式で書き込みが先に起こった場合を考える と，読み出しのスレッドを最初からやり直す必要が出てくる．そこで，書き込みが先の場合は主記憶の値を変更し，他のスレッドが書き込む場合などに備えてやり直し用のデータを残しておく方式とした．

「マルチコアが変えるソフトウェア開発」，日経エレクトロニクス 2006.10.09, No. 936, p.67--87 より引用

[他の参考文献]

Maurice Herlihy, et al.,Transactional Memory : Architectural Support for Lock-Free Data Structures, ISCA, pp.289--300, 1993.

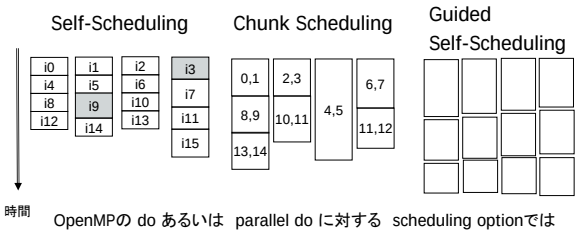
Tim Harries, et al., Transactional Memory : An Overview, IEEE Micro, Vol.27, No. 3, pp.8-29, May/June 2007.

負荷の不均衡の解消

- 動的負荷分散方式
  - Doall型の例
    - Self-Scheduling，Chunk Scheduling，Guided Self-Scheduling
  - 反復計算向けの負荷分散アルゴリズム
    - N回目の計算時の実行時情報からN+1回目の計算時の最適な処理の分割を予測する。
      ⇒ 非均質な計算環境への適用可(heteroTINPAR by 富田研)
  - OSレベルでの対応
    - Cache Affinity Scheduling (ラストプロセッサ方式、最小介入数方式，等)
    - Memory Affinity Scheduling
      - 1)ホーム，2)コピー， 3)Minimum Load Cluster，4)Other
    - ギャングスケジューリング/Co-scheduling
  - その他……HTGを利用した実行時粒度制御機構

負荷の不均衡の解消

- (広義の)Self-scheduling方式

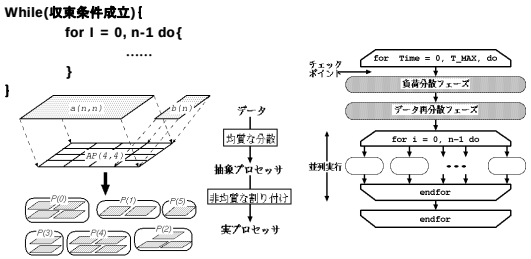


OpenMPの do あるいは dynamic(1),

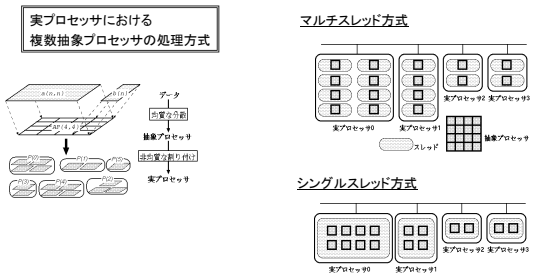
parallel do に対する dynamic(2)

scheduling optionでは guided(n) に相当

非均質環境における動的負荷分散の例 (heteroTINPAR)



非均質環境における動的負荷分散の例 (heteroTINPAR)



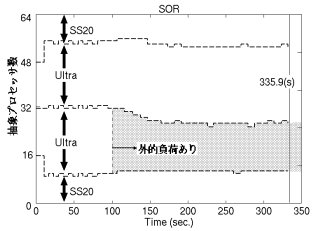
非均質環境における動的負荷分散の例 (heteroTINPAR)

動的負荷分散による実行時間の変化

評価環境： SS20x2台＋Ultra1 x 2台
評価プログラム：
SOR(1024x1024, 2000反復)
抽象プロセッサ数：64

実行開始100秒後に一方のUltra1に負荷をかける

負荷分散なしの実行時間＝357.6(s)



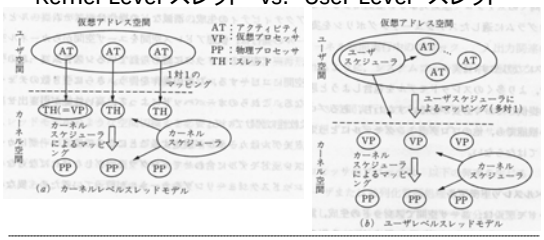
動的負荷分散システムの分類

- 同期Synchronizationの観点での分類

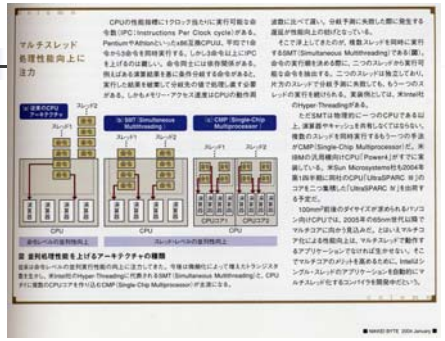
(Loosely) Synchronous	Stop-and-Repartition (Explicit)
Asynchronous	Poll-driven Load Balancing (Explicit)
	Interrupt-driven Load Balancing (Implicit)

マルチスレッド環境

Kernel Level スレッド vs. User Level スレッド



[HW実装(余談)] Simultaneous Multi-threading(SMT) …Intel用語ではHT



SSIクラスタ

- Complete Transparency…. Shared Address Space
- Scalable Performance……..Small SSI Overhead
- Enhanced Availability……. Fault Tolerant

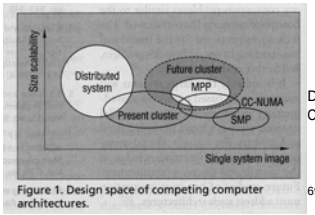
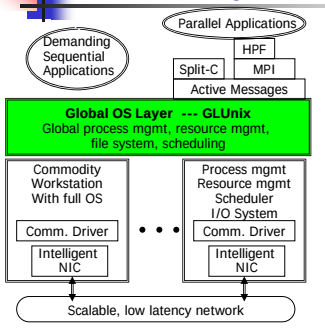


Figure 1. Design space of competing computer architectures.

Designing SSI Clusters with Hierarchical Checkpointing and Single I/O Space

Kai Hwang, Hai Jin, Edward Chow, Cho-Li Wang, and Zhiwei Xu
IEEE Concurrency, Vol.7, No.7, pp.60—69, January-March 1999

UC Berkley “NOW”

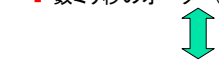


GLUnix
UNIX のシステムコールを横取り
SSI環境を実現

xFS ….Serverless FS
Network RAID(Striping)
Cooperating Caching
1)Clientの主記憶をFile Cacheとして使用
2)MP用のCC機構を利用してFileを管理

近くのハードディスク vs 遠くのメモリ

- HDDの転送速度
  - SCSI系 ～ 320MB/s(ULTRA SCSI320)
  - Fibre Channel 1.06～4.24Gbps
  - IEEE1394系 ～ 400Mbps
- HDDのレイテンシ
  - 数ミリ秒のオーダ（DISKの回転速度が支配）



- ネットワークの転送速度 ……. 1～10Gbps
- ネットワークのレイテンシ……. 数十マイクロ秒