

コンピュータシステム（第7回）

坂井 修一

東京大学大学院 情報理工学系研究科 電子情報学専攻

- ・ はじめに
- ・ 相互結合網

コンピュータシステム

東大・坂井

はじめに

- 講義内容： コンピュータシステム
 - － 並列処理、省電力、ディペンダビリティ、コンピュータの未来
- 関連講義（電子・情報系）：
 - － デジタル回路：入江、2年冬
 - － コンピュータアーキテクチャ：坂井、3年夏
 - － アドバンスト・コンピュータアーキテクチャ：入江、大学院奇数年度夏
- 成績
 - － 出席＋レポート

コンピュータシステム

東大・坂井

相互結合網（Interconnection Network）

- 役割
 - － プロセッサ間通信、プロセッサ・メモリ間通信
 - ・ 高スループット
 - ・ 小レーテンシ
 - ・ 実時間性の保持
 - ・ 低コスト
 - － 付加機能
- 歴史： 非常に古い！
 - － マルチプロセッサの進歩の大きな一部

コンピュータシステム

東大・坂井

相互結合網の研究開発

- トポロジー
- 交換方式
- ルーティング
- 仮想チャネル
- デッドロック対策／ライブロック対策
- 付加機能
- 実装

コンピュータシステム

東大・坂井

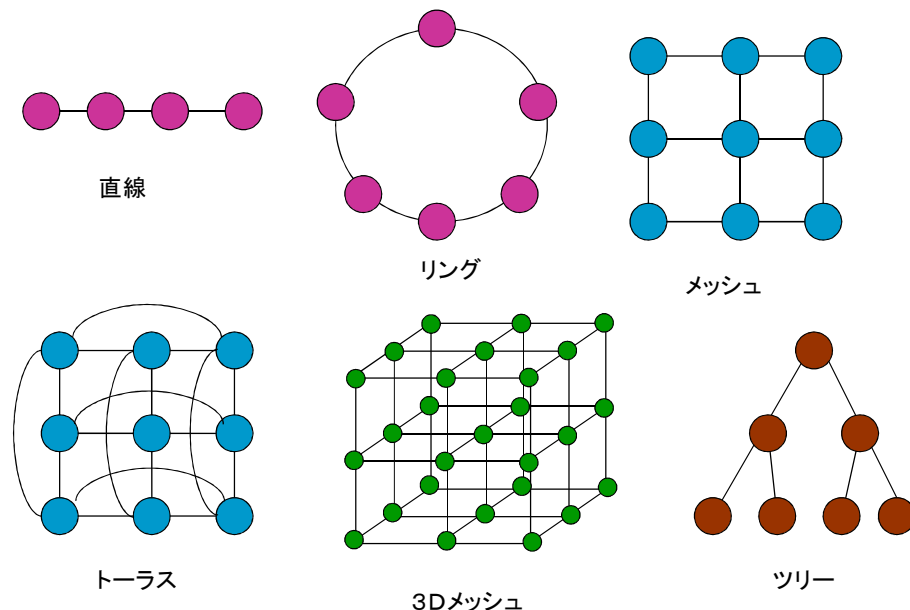
相互結合網のトポロジー

- 直接網: ノードにデータ交換機能が入っている
 - グラフの直径が $O(N)$: 直線、リング
 - $O(N^{1/2})$: メッシュ、トーラス
 - $O(N^{1/3})$: 3Dメッシュ、3D トーラス
 - $O(\log N)$: ツリー、超立方体(ハイパーキューブ)、De Bruijn, CCC, サーキュラオメガ、CCCB
 - $O(1)$: スター、完全結合
- 間接網: ノードの外にデータ交換用のスイッチがある
 - 段数1: クロスバスイッチ
 - 3: クロス(Clos)
 - $\log_M N$: ハイパクロスバ (Mは各次元あたりのノード数)
 - $\log N$: オメガ、ベースライン、デルタ、間接キューブ、データマニピュレータ
 - $2\log N - 1$: ベネス、ファットツリー
 - $(1/2)\log N(\log N + 1)$: バイトニックソート

コンピュータシステム

東大・坂井

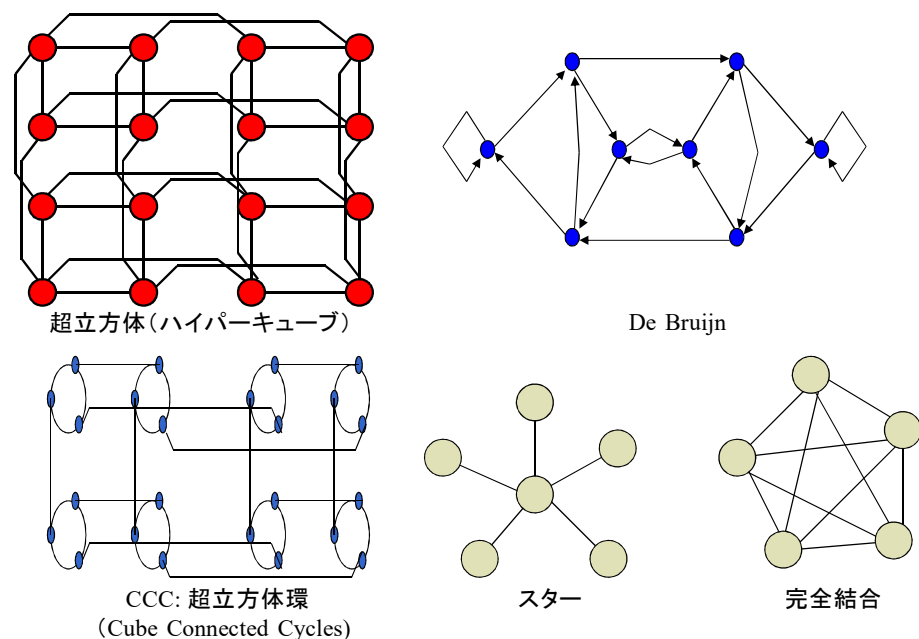
直接網の例



コンピュータシステム

東大・坂井

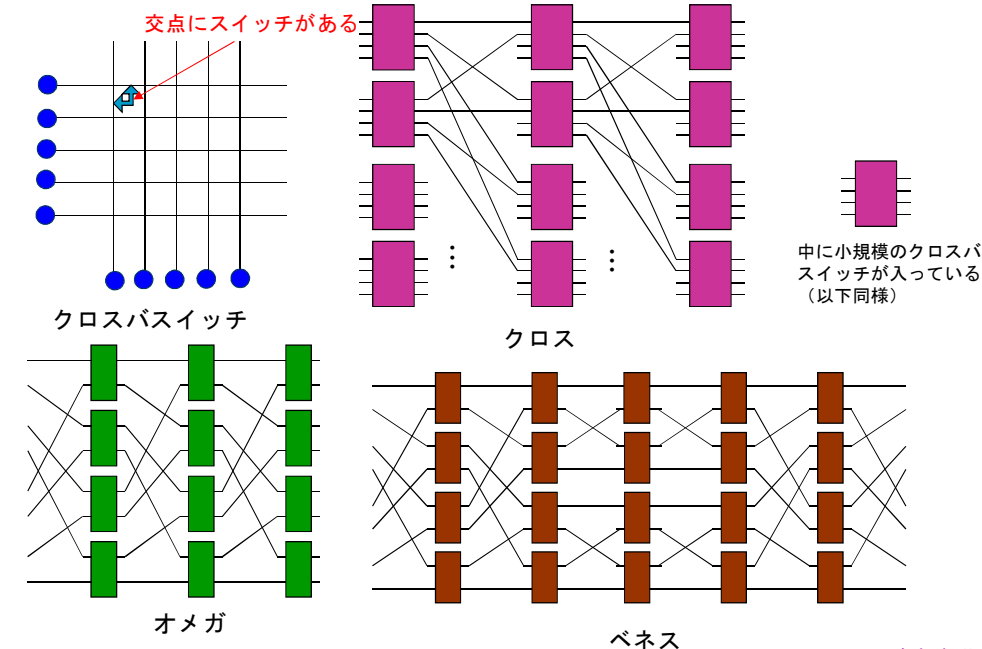
直接網の例(続き)



コンピュータシステム

東大・坂井

間接網の例



コンピュータシステム

東大・坂井

ネットワークの評価

コスト

- 交点の数(C)
- ノード(またはスイッチ)のポート数(P)
- 線の数(L)

パフォーマンス(1): 静的な指標

- ノード間の距離(グラフの直径、D)
- バンド幅(B_S): 解析で求まるもの
- 二分バンド幅: (bisection bandwidth、BB)

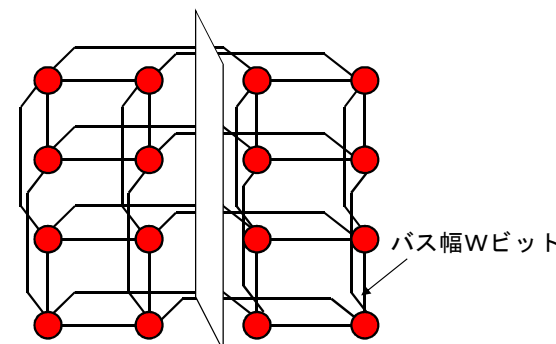
パフォーマンス(2): 動的な指標

- バンド幅(B_D): シミュレーション、実装で求まるもの
- 遅延時間(D): 同前

ネットワークの評価(続き)

2分バンド幅 (Bisection Bandwidth、BB)

- 並列システムを2つの部分に分割したときに、カットする線の数の最小値



- 2分バンド幅が大きい
 - 大域的な通信に強い
 - 配線が大変
- 2分バンド幅が小さい
 - 大域的な通信に弱い
 - 配線がやりやすい

ハイパーキューブの場合 $BB = W * N / 2$

ネットワークの評価(続き)

ネットワーク	交点数 C	ポート数 P	線数 L	直径 D	二分バンド幅BB
直線	$9(N-2)+8$	6	$W(N-1)$	$N-1$	W
リング	$9N$	6	WN	$N/2$	$2W$
メッシュ	$25N-36N^{1/2}+8$	10	$2WN^{1/2}(N^{1/2}-1)$	$(N^{1/2}-1)*2$	$WN^{1/2}$
トーラス	$25N$	10	$2WN$	$(N^{1/2}-1)$	$2WN^{1/2}$
3D メッシュ	$49N-6N^{2/3}-264N^{1/3}+288$	14	$W(3N-6N^{2/3}+3N^{1/3}+6)$	$(N^{1/3}-1)*3$	$WN^{2/3}$
3D トーラス	$49N$	14	$3WN$	$(N^{1/3}-1)*(3/2)$	$2WN^{2/3}$
超立方体	$N(\log N+1)^2$	$2\log N+2$	$(W/2)(N\log N)$	$\log N$	$WN/2$
ツリー(2進)	$10N-13$	8	$W(N-1)$	$2\log(N+1)-2$	W
スター	N^2+4N-4	$2N$	$W(N-1)$	2	$WN/2$
完全結合	N^2	$2N$	$WN(N-1)/2$	1	$WN^2/4$
クロスバ	N^2	-	WN^2	1	$WN^2/2$
ハイバクロスバ (d次元)	$dN^{(d+1)/d}$	-	$dWN^{(d+1)/d}$	d	$WN^2/2$
オメガ	$2N\log N$	-	$WN(\log N+1)$	$\log N+1$	WN
ベネス	$2N(2\log N-1)$	-	$2WN\log N$	$2\log N$	$2WN$

ネットワークの評価(続き)

直接網

- 距離 $\log N$ (場合により $N^{1/d}$), コスト $O(N)$ が
- 超立方体は、コストがスケラブルでない！
- シャフル系、メッシュ系などが中心
 - ・ シャフル系: De Bruijn, CCC, CCCB
 - ・ メッシュ系: 3Dメッシュ、階層メッシュ

間接網

- 実用的には $\log N$ 段の網が有望
 - ・ ルーティングが簡単(後述)
 - ・ ブロッキング(回線の干渉)が問題
- ベネス網はノンブロッキング: ただし、rearrangeable
 - ・ 一本のつながかえに、全体の変更が必要
- クロス以上は、ハード量が大きい

交換方式

■ 回線交換

- 送り手と受け手の間に回線を設定してから、送りたい情報全部を転送する

■ パケット交換

- パケットと呼ばれる小さな単位によって、細切れにした情報を送る



パケット交換方式

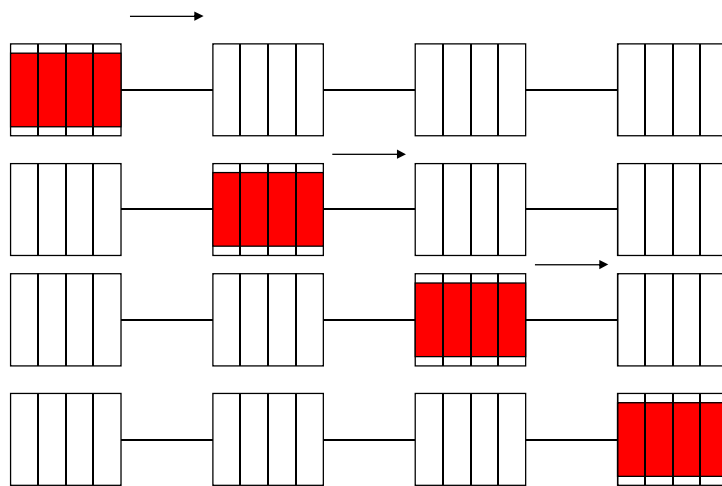
- ストアアンドフォワード
 - ・ 中継ノードごとにパケット全体をバッファリングする
- ワームホール
 - ・ 回線が空いていれば、パケットは次々に進める
 - ・ 回線がふさがっていると、パケットの転送全体が止まる
- バーチャルカットスルー
 - ・ 回線が空いていれば、パケットは次々に進める
 - ・ 回線がふさがっていると、パケットの転送はそこでバッファに溜められる

- ハードウェア資源が豊富なときは、一般にバーチャルカットスルー方式が優れている

コンピュータシステム

東大・坂井

ストアアンドフォワード方式

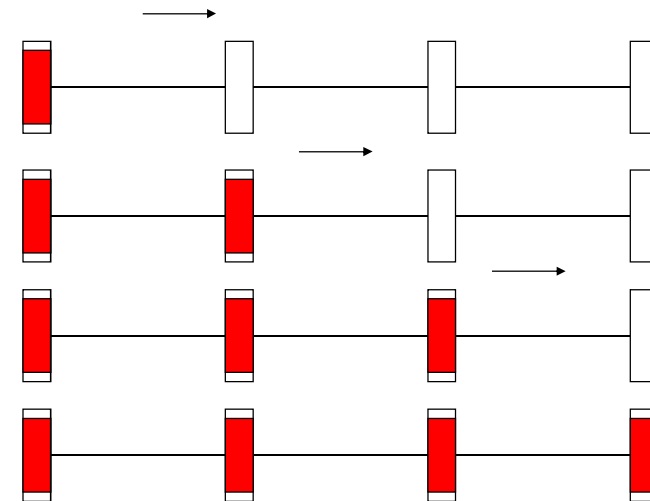


- 遅延: $T * P$, T は一段あたり1ワードを転送する時間、 P はパケット長
- ソフトウェアで制御可能
- バッファ量は、パケットの最大長以上必要

コンピュータシステム

東大・坂井

ワームホール方式

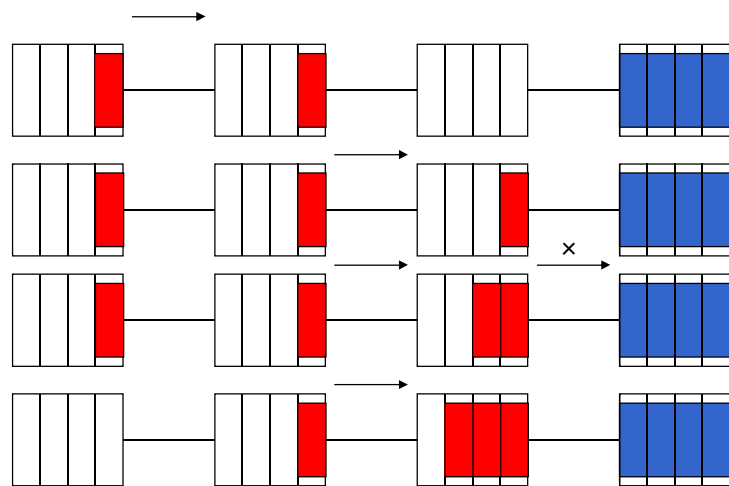


- 遅延: $T * H + D$, T は一段あたり1ワードを転送する時間、 H はヘッダ長、 D は(ペイロード+テイラ)長
- バッファ量は、1語分だけでよい
- 専用ハードウェアが必要となる

コンピュータシステム

東大・坂井

バーチャルカットスルー方式



- 遅延 : $T * H + D$, T は一段あたり 1 ワードを転送する時間、 H はヘッダ長、 D は (ペイロード+テイラー) 長
- バッファ量は、2 語以上必要な大きさだけ
- 専用ハードウェアが必要となる

コンピュータシステム

東大・坂井

ルーティング

- 集中型と分散型
 - 分散型が主流に
- ルーティングアルゴリズム
 - 行先アドレスの1ビット参照(オメガ)
 - マンハッタンルーティング(メッシュ)、など
- 固定ルーティングと可変ルーティング
 - 固定ルーティング
 - 可変ルーティング
 - ・ 故障の迂回
 - ・ 適応型ルーティング
 - 混雑の回避

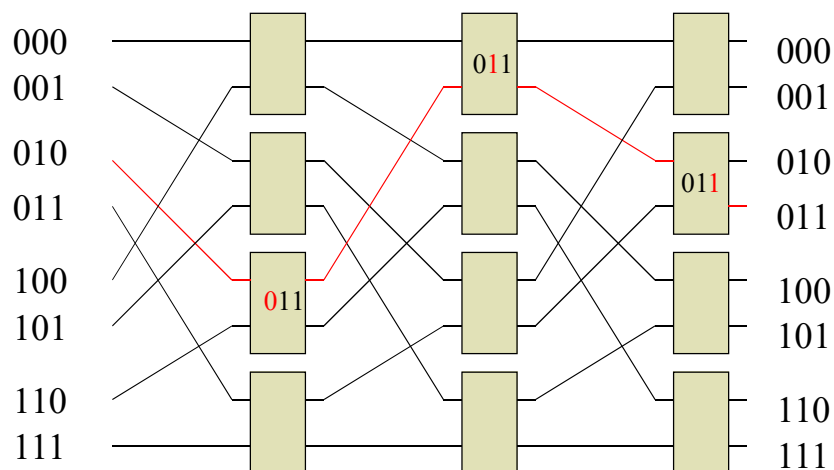
コンピュータシステム

東大・坂井

1ビットルーティング(オメガ網)

- 行先の番地を上の桁からチェックし、0ならば上、1ならば下に進む

010 → 011 へのルーティング



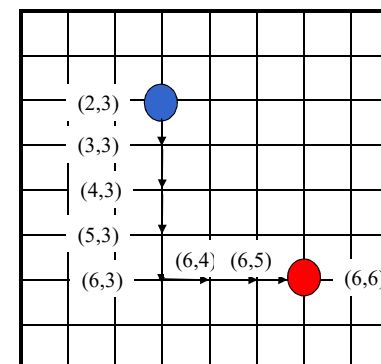
コンピュータシステム

東大・坂井

マンハッタンルーティング(メッシュ)

- まず南北方向に進めて緯度を合わせ、続いて東西方向に進めて目的地に到達する(南北と東西の順序は逆でもよい)

(2,3) → (6,6) へのルーティング



マンハッタンルーティングの利点 : ストアアンドフォワードデッドロック (次項参照) が生じない

コンピュータシステム

東大・坂井

デッドロックとライブロック

■ デッドロック

- ストアアンドフォワードデッドロック (SFD)
 - 転送要求が網上でループを成し、ループ上のバッファがすべてフルで、パケットが動けなくなった状態
 - ・ 直接SFD: 2ノード間でデッドロックが生じる
 - ・ 間接SFD: 3ノード以上でループが発生する
- リアセンブリデッドロック
 - 行先でのバッファ量不足により、パケットからもとのデータを再構築(reassembly)できなくなった状態
- リシーケンスデッドロック
 - 行先でのバッファ量不足により、データをもとの順番に並べ替える (resequence)できなくなった状態

■ ライブロック

- ある通信が、他の通信によってブロックされていていつまでたっても実現されない状態
- 公平性(Fairness)の保証によって回避される

コンピュータシステム

東大・坂井

SFD防止

■ SFDの防止

- グラフからループを取り除けばよい

■ (例)メッシュにおけるSFD防止

- マンハッタンルーティング(前述)

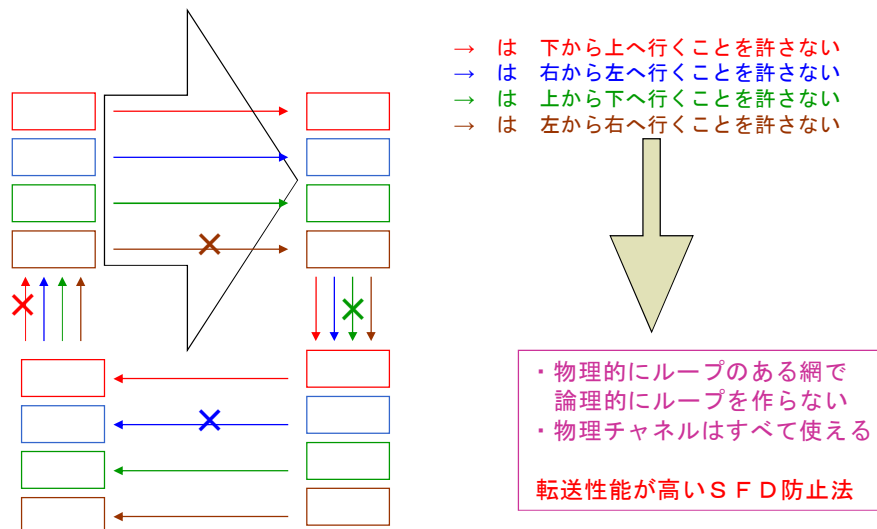
■ 一般的なSFD防止法

- 構造化バッファ法
 - ・ 直径の数のバッファを各ノードに用意する
 - ・ 1回ホップするたびに一つ上位のバッファを使う
 - ・ 最上位のバッファに到達する前に行先に到達するから、ループは発生しない
 - ・ 欠点: 最低でも直径の大きさのバッファが必要
- バッファ分割+仮想チャネル
- ・ 網の上に複数の論理的なネットワークを構築し、それぞれにループが生じないようにする

コンピュータシステム

東大・坂井

バッファ分割+仮想チャネルによるSFDの防止

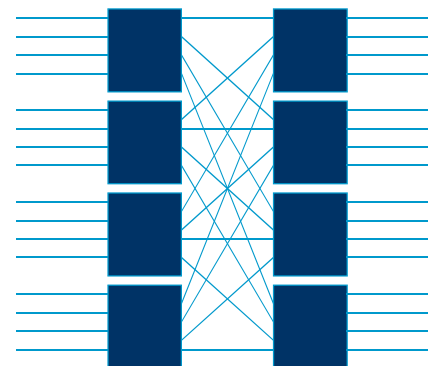


コンピュータシステム

東大・坂井

実装例(1)

■ PIE64 : Omega Network



- ・ 4入力ポート、4出力ポート
- ・ バイチャルカットスルー型パケット交換
- ・ 負荷のバックワード伝播による自動負荷分散
- ・ デバッグ機構



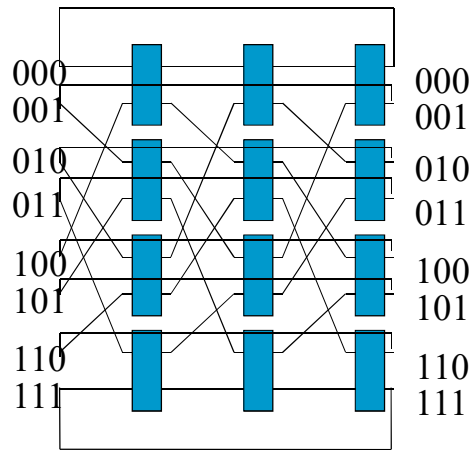
PIE64の実装

コンピュータシステム

東大・坂井

実装例(2)

■ EM-4: Circular Omega Network



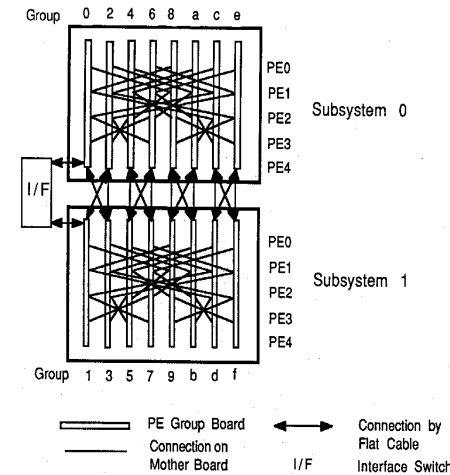
- 3入力ポート、3出力ポート
- $O(\log N)$ の距離
- バイチャルカットスルー型パケット交換
- 細粒度パケット処理
 - パケット長: 80bit 固定長
 - 1クロックパケット生成
 - ユーザレベル0コピー通信
 - 細粒度同期
- 仮想チャネルによるSFD防止
- 裏循環路による自動負荷分散機能
- 実装
 - シングルチッププロセッサの一部
 - 約61 MB/s/port
 - 80ノードを実装

サーキュラオメガ網: オメガ網のスイッチ部にプロセッサ+スイッチをつけた網

コンピュータシステム

東大・坂井

EM-4ネットワークの実装



EM-4網構成: 80ノードを実装

コンピュータシステム

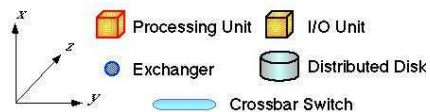
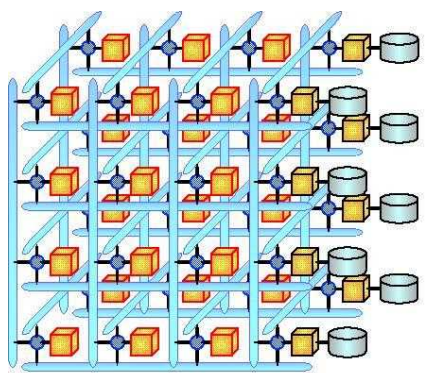


EM-4プロトタイプ: 60cm*92cm*140.5cm

東大・坂井

実装例(3)

■ CP-PACS(SR2201): Hypercrossbar



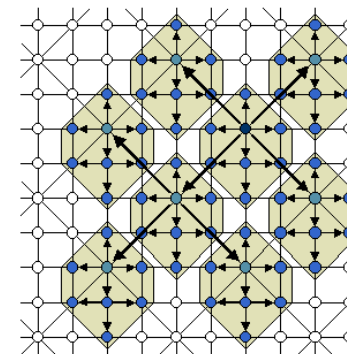
- 3階層の構造
 - ボード: 8×8 クロスバ
 - ラック: 17×17 クロスバ
 - 全体: 16×16 クロスバ
- ワームホール型パケット交換
- ユーザレベルゼロコピー通信
 - Remote DMA
 - 300 MB/s/port

コンピュータシステム

東大・坂井

実装例(4)

■ Jump-1: RDT (Recursive Diagonal Torus)



- 直径が大きいトーラス網の欠点を階層化により緩和
 - 1階層: 普通のトーラス
 - 2階層: 1階層の対角線を使ったトーラス
 - 3階層: 2階層の対角線を使ったトーラス
- 共有メモリ支援機構: 階層化ビットマップ
- マルチキャスト機能

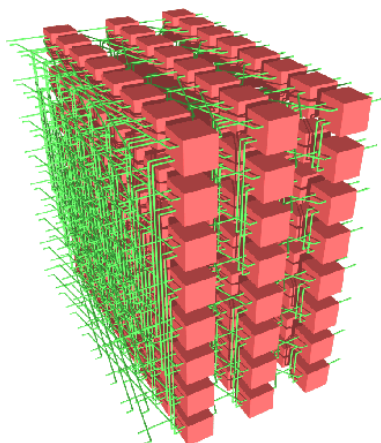
コンピュータシステム

東大・坂井

実装例(5)

■ RWC-1:

CCCB (Cube Connected Circular Banyan)



- 4 入力ポート、4 出力ポート
- $O(\log N)$ の距離
- バイナルカットスルー型パケット交換
- 仮想チャネルによる SFD 防止
- マルチプロセス支援
 - パーティショニング
 - ドレイン処理
- 仮想制御チャネルによるカーネルパケット処理
- 2 本の裏循環路による負荷分散
- 実装
 - BiCMOS Switch Chip: 12 bit bit slice * 2
 - 300 MB/s/port

IEEE Outstanding Paper Award受賞

コンピュータシステム

東大・坂井

RWC-1ネットワークの実装



RWIC

Tsukuba Research Center, Real World Computing Partnership



1 筐体 64 プロセッサ (右はRAID)
128 プロセッサ版を実装・稼動

コンピュータシステム

東大・坂井

これからの相互結合網

- コモディティの活用
 - Tera bit Ethernet
 - WDM化による高帯域化
- HPCネットワークとPCクラスタネットワーク
 - カスタム部分の設定
- ソフトウェア
- ディペンダビリティ
 - 信頼性
 - 安全性
 - 拡張性
 - 堅牢性

コンピュータシステム

東大・坂井