

## データ工学特論(5)

情報基盤センター  
天野 浩文

この講義に関するwebサイト:

<http://isabelle.cc.kyushu-u.ac.jp/~amano/data-engineering/>

1

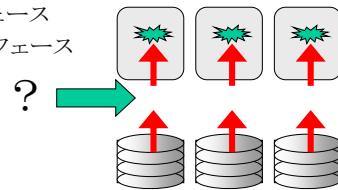
### 前回のおさらい(1)

- 並列プログラムのデータ入出力
  - ◆ 単一プロセッサに代表してやらせる
  - ◆ 各プロセッサで並列にやる
- 並列入出力のためのしくみ
  - ◆ 並列ファイルシステム
    - Vesta
    - Acacia
- 当初の予定では、この後に...
  - ◆ 並列入出力ライブラリ
  - ◆ 並列入出力インタフェース
  - ◆ ...と進むはずだったが、今回は少し予定を変更する.
- 次回の特別講義の下準備のため

2

### 前回までに扱ってきたことは...

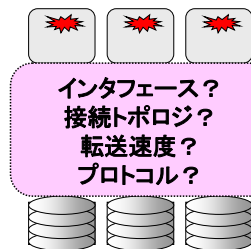
- これまでの講義では
  - ◆ 転送速度・スループットを向上させるために  
⇒ **ディスク側の速度向上**
    - ディスクアレイ, RAID (redundant array of inexpensive disks)
  - ◆ 並列プログラムへのデータ供給を効率化するために  
⇒ **並列プログラムの並列データアクセス性能の強化**
    - 並列ファイルシステム
    - 並列入出力インタフェース
    - 集団型入出力インタフェース



3

### さて、今回からは...(1)

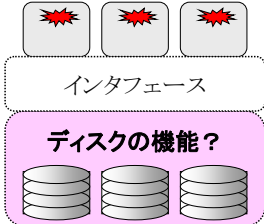
- 現代の計算機システムにおけるディスクのつなぎ方はどのように変わりつつあるか?
- 計算機とディスク(アレイ)の間の接続方法は?
  - ◆ インタフェース
  - ◆ 接続トポロジ
  - ◆ 転送速度
  - ◆ プロトコル
- 各種ディスク入出力インタフェース
  - ◆ SCSI
  - ◆ HIPPI
  - ◆ Fibre Channel
  - ◆ GSN (Super HIPPI, HIPPI-6400)



4

### さて、今回からは...(2)

- ディスクの機能は今のままで十分か？
  - ◆ホスト側のCPUにかかる負担は？
- ホストに負担をかけないためには？
  - ◆SAN (storage area network)
  - ◆NAS (network attached storage)
- さらに賢いディスク装置に向けて
  - ◆高性能ディスク
  - ◆自律ディスク



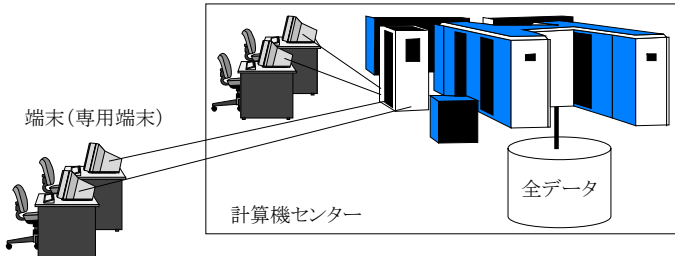
インタフェース

ディスクの機能?

5

### データ配置から見た計算機利用形態の変遷(1)

- 集中処理の時代(1970年代～1980年代半ば)
  - ◆1台のホスト(メインフレーム)による集中処理
  - ◆ディスク(データ)はすべてホストに集中
  - ◆利用者は、端末から専用線経由でホストにアクセス(端末側では、独自の処理をほとんど行わない)



専用端末

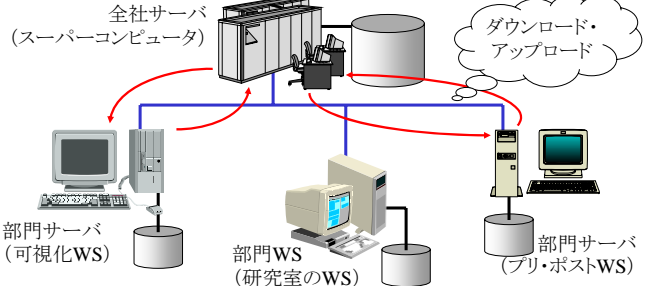
計算機センター

全データ

6

### データ配置から見た計算機利用形態の変遷(2)

- 分散処理の時代(1980年代～)
  - ◆LAN, ワークステーションの台頭
  - ◆組織全体のサーバだけでなく各部門のサーバにある程度のデータを移して、分散処理を行うことが可能に



全社サーバ (スーパーコンピュータ)

ダウンロード・アップロード

部門サーバ (可視化WS)

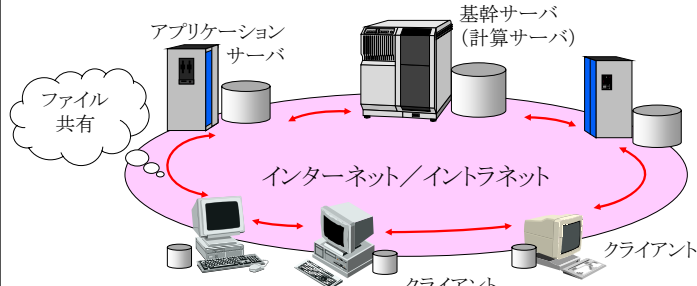
部門WS (研究室のWS)

部門サーバ (アプリ・ポストWS)

7

### データ配置から見た計算機利用形態の変遷(3)

- 協調処理の時代へ(1990年代後半～)
  - ◆インターネット/イントラネットの普及
  - ◆データのさらなる大容量化
  - ◆複数のサーバ間でのデータの共有が急速に進行



アプリケーションサーバ

基幹サーバ (計算サーバ)

インターネット/イントラネット

クライアント

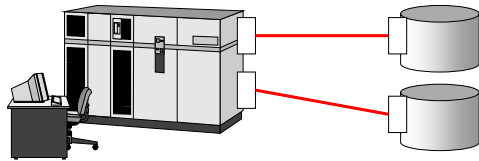
クライアント

ファイル共有

8

### ディスク接続形態の変遷(1)

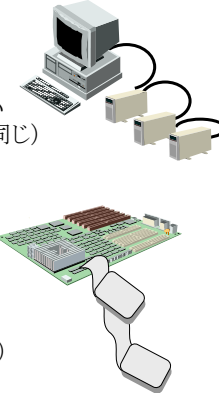
- 1対1接続
  - ◆各ディスク装置に対して、専用のチャンネルを用意
    - ・主としてメインフレーム用に発達
  - ◆各ベンダで固有の仕様
    - ・電気チャンネル
    - ・光チャンネル
  - ◆最大ケーブル長は数十～数千m



9

### ディスク接続形態の変遷(2)

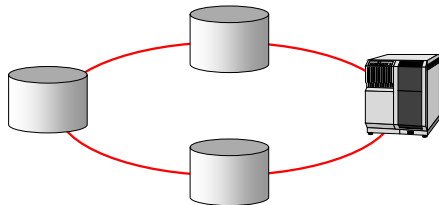
- デジチェーン接続(1対1接続の一種と見ることもできる)
  - ◆1つのインタフェースに複数のディスクを数珠つなぎ
  - ◆やや拡張性が増す
  - ◆複数ディスクを同時に使うことはできない(転送速度,スループットは1台のとときと同じ)
  - ◆例1:SCSI (small computer systems interface)
    - ・インタフェースあたり7～15台
    - ・最大ケーブル長(合計)は1.5～3m
  - ◆例2:IDE (integrated drive electronics)
    - ・インタフェースあたり2台まで
    - ・ケーブル長は, PCのケース内(~1m)



10

### ディスク接続形態の変遷(3)

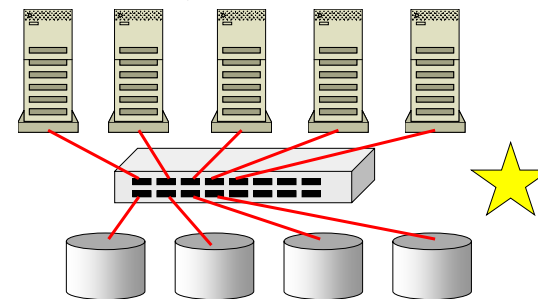
- ループ接続
  - ◆ホスト, ディスクともに, それぞれ2つの通信口を持つ
  - ◆それらをループ状に接続する
  - ◆自分が処理に参加していないときは, 他の通信をそのまま通す
  - ◆複数ディスクを同時に使うことはできない(最大転送速度,スループットは1台のとときと同じ)



11

### ディスク接続形態の変遷(4)

- スター型接続, (ファブリック接続, スイッチ接続)
  - ◆スイッチやハブを通して, 星状に接続
  - ◆複数ディスクを同時に使うことも可能(スループットが向上)



12

### ディスク入出インタフェース(1)

- SCSI (Small Computer Systems Interface) 規格
  - ◆基本的に、単一の計算機内での周辺機器の接続を想定
  - ◆デイジーチェーン接続
  - ◆接続ケーブルの電氣的仕様から、CPU側と周辺機器側でのデータのやりとりに用いるSCSIコマンドプロトコルまでを規定
  - ◆IDEと異なり、ディスクコントローラ側で処理を分担するため、CPU負荷がやや軽くなる
  - ◆ただし、初期のSCSIは高速ではなかった
    - SCSI (5MB/sec), SCSI-2 (10MB/sec), SCSI-3 (20MB/sec)
- 高速のSCSIも登場
  - ◆Ultra2 SCSI: 40MB/sec
  - ◆Ultra Wide SCSI: 80MB/sec
  - ◆Ultra160 SCSI: 160MB/sec

13

### SCSIの現状

- 個人用のPCではほとんど使われなくなった。
  - ◆ディスクドライブが大容量になり、IDEのディスク1台～2台で十分な容量が確保できる。価格も安い。
  - ◆増設が必要な場合でも、シリアル系のUSBやIEEE1394など、高速のインタフェースで簡単に外付けできるようになった。
  - ◆CPUの性能が上がったため、ディスク制御の負荷がかかっても気にならなくなった。
- SCSIケーブルは信号線が複数並列にたばねられたパラレルケーブルだった。
  - ◆信号を同期させるための回路や、クロストーク対策などが必要
  - ◆最近では、シリアルケーブルで周波数を上げるほうが高速化しやすい。
  - ◆次世代 SCSI 規格は serial attached SCSI (SAS)に移行中

14

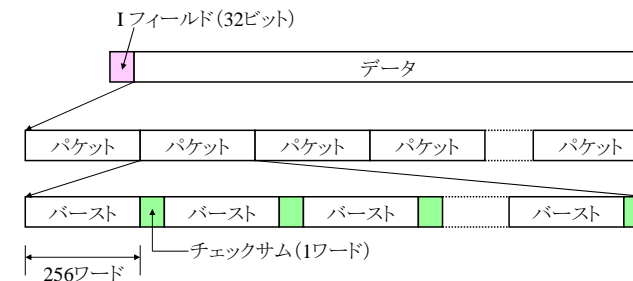
### ディスク入出インタフェース(2)

- HIPPI (high-performance parallel interface)
  - ◆元々、スーパーコンピュータと周辺機器を接続するためのインタフェースとして登場
    - その後、計算機間の接続にも
  - ◆銅線ケーブルで25m以内
    - 100MB/sec (32bit幅, 50芯ケーブル使用)または200MB/sec (64bit幅, 100芯ケーブル使用)
    - その後、ファイバ系 (Serial HIPPI)も登場
  - ◆1本のケーブルは単方向通信用, 双方向通信にはケーブルをペアで使用
  - ◆1対1接続またはHIPPIスイッチ経由でスター型接続
  - ◆遅延は大きいので、大量データを一気に転送する用途向け
  - ◆(スパコン専用のため)市場規模小さく、価格下がらず

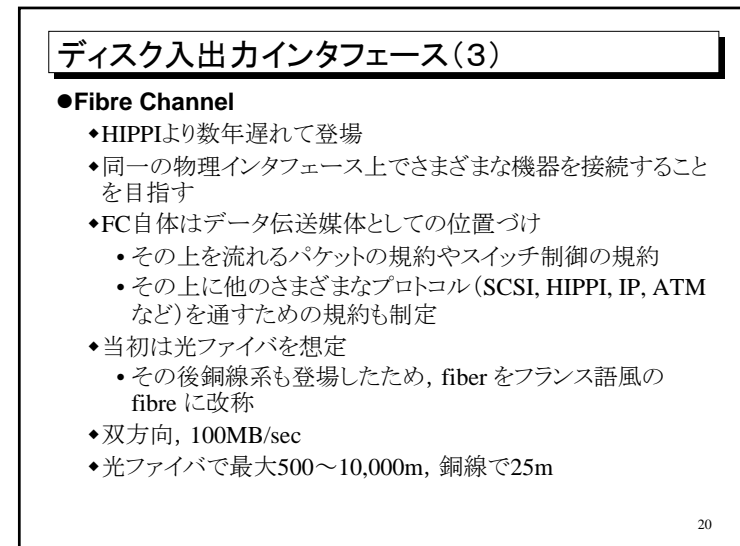
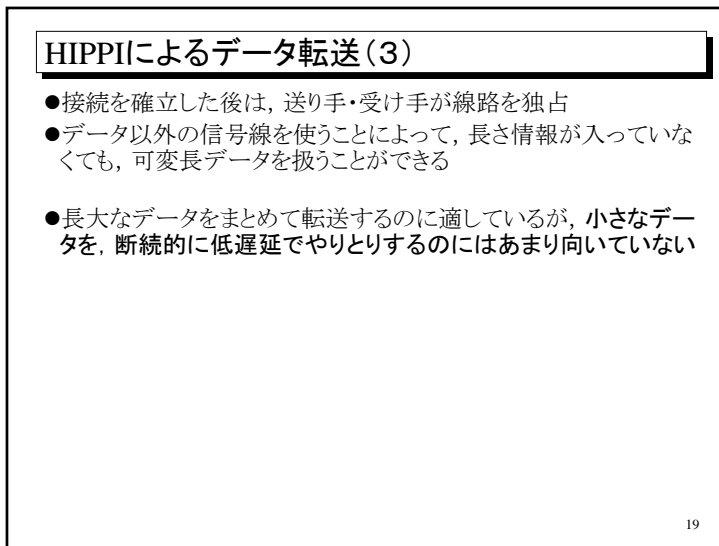
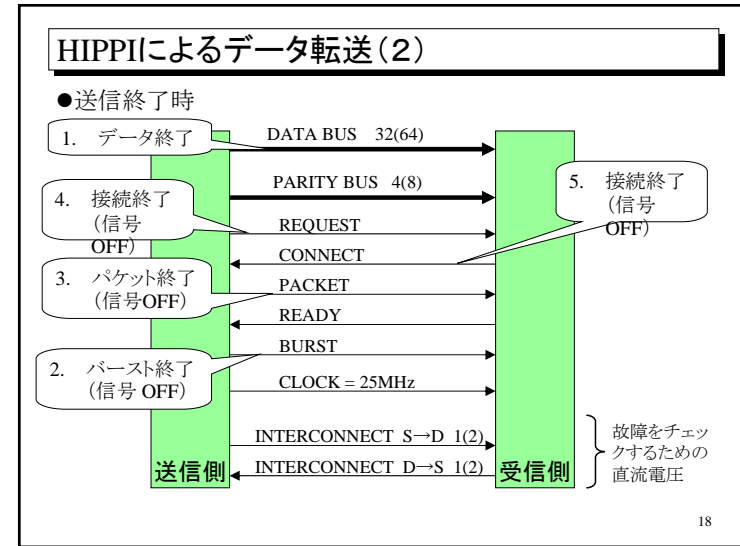
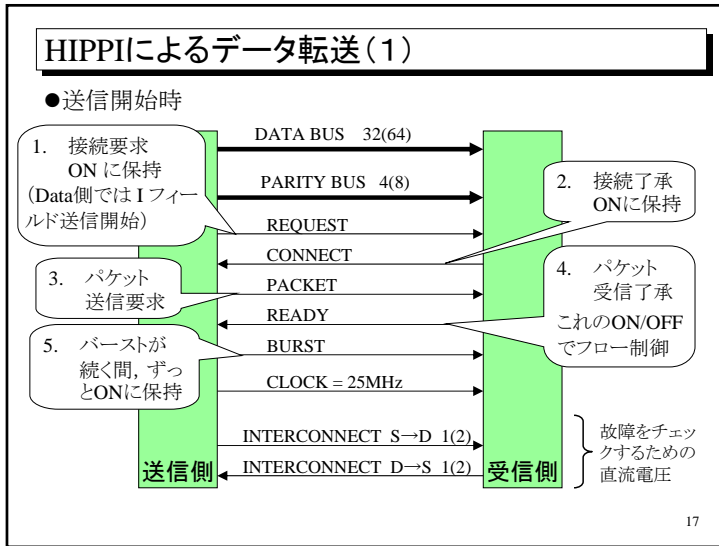
15

### HIPPIフレーム

- 送受信の単位は256ワード(1ワード32ビットまたは64ビット)のバースト
- 先頭のヘッダ情報(Iフィールド)は、スイッチの経路制御用
- 接続を確立した後は、その送り手・受け手が線路を独占
- フレームの長さは約4GBまで可(データには長さの情報は含まれていない)



16



### FCデータフレーム

- フレームごとに、送り手・受け手のアドレスその他の情報がヘッダとして付随
- 1フレームあたりの最大データ量は2112バイト
- 1フレームに収まり切れない場合は、複数のフレームに分割  
→シーケンス

21

### ファイバチャネルの特徴

- すべての通信は、比較的小さなフレームにバラして送受信する
- すべてのフレームに送り手・受け手のアドレスがついている

↓

- 途中のスイッチが高速に配送経路の切り替えを行えば、複数の通信を同じ線路上で同時に流せる

- HIPPIに比べると安価で、現在急速に普及しつつある

22

### HIPPI vs. Fibre Channel

	HIPPI	Fibre Channel
最大転送速度 (Mbps)	800, 1600	100, 200, 400, 800 1600, 3200
主な伝送媒体	銅ケーブル	光ファイバ
伝送形態	並列	直列
通信形態	独占(多重化不能)	多重化可能
オプション	転送速度のみ	多種多様
接続形態	1対1, スター	1対1, スター, ループ他
製品化	1988年	1993年
欠点	高価 低速の通信を多数扱うことはできない	プロトコルが複雑すぎるとの批判も

23

### Fibre Channel による接続の形態 (1)

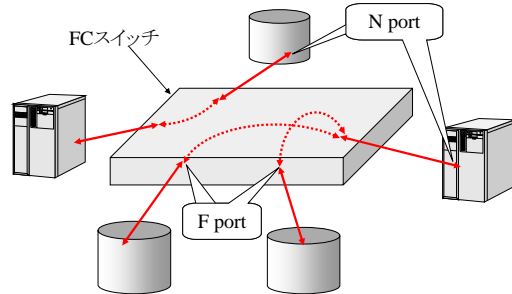
- Point-to-Point型 (2ノード)
  - ◆N (node) ポートどうしを接続

- Fibre Channel-Arbitrated Loop (FC-AL, 最大127ノード)
  - ◆ループ機能を持つNL (node loop) ポートを使用

24

### Fibre Channel による接続の形態(2)

- Fabric型(スイッチのポート数次第で最大1600万ノードまで)
  - ◆任意のNポート間を自由に接続できるFCスイッチを使用
  - ◆スイッチ側はF (fabric) port



25

### ディスク入出カウンタフェース(4)

- GSN (Gigabyte System Network)
  - ◆次世代HIPPIを目指す (Super-HIPPI, HIPPI-6400)
  - ◆6.4 Gbps (800MB/sec)
  - ◆HIPPI-800の欠点を克服することを目指す
    - 低遅延(通信の立ち上がりが速いということ)
    - マイクロパケット(32バイト+制御用64ビット)単位
    - 仮想チャネルによる多重化(大容量データの転送中にも、スイッチ制御などのためのメッセージを割り込ませることができる)
    - 豊富な機能を盛り込む
      - エラー訂正, フロー制御, 再送制御
    - ただし, 価格の問題は?
  - ◆ようやく製品が出始めた
  - ◆普及はまだまだこれから?

26

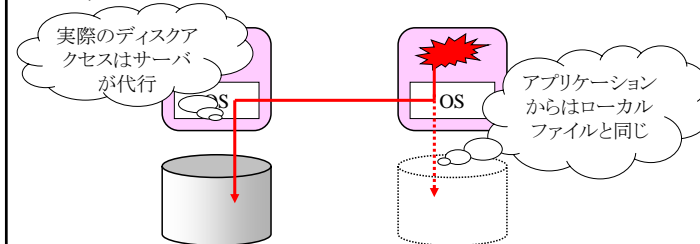
### Scheduled Transfer (ST) プロトコル

- GSN上の伝送プロトコル
  - ◆フロー制御, エラー制御, バッファ管理, 遠隔DMA転送などの機能を実現
  - ◆他のメディア上でも利用可能
- マイクロパケット上で高速・大容量転送を実現する
  - ◆接続の確立
  - ◆受け手は転送のために連続領域のメモリバッファを確保
  - ◆以後は, メモリ→メモリの転送をホスト側プロセッサの介入なしに実行
- イーサネットでのCPU負荷を100%とすると, STでは15%ほどと言われる
- この上にSCSIプロトコルを載せる技術が開発中

27

### ファイル共有の形態

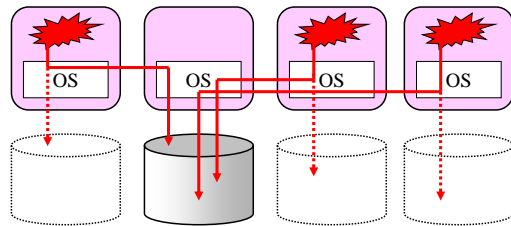
- NFS (Network File System), CIFS (Common Internet File System), SMB (Server Message Block, "Samba")
  - ◆他のホストにあるファイルをあたかもローカルファイルのようにアクセスするための機構
  - ◆クライアントからのリクエストに応じて, サーバがディスクアクセス



28

### ネットワーク上のファイル共有の問題点(1)

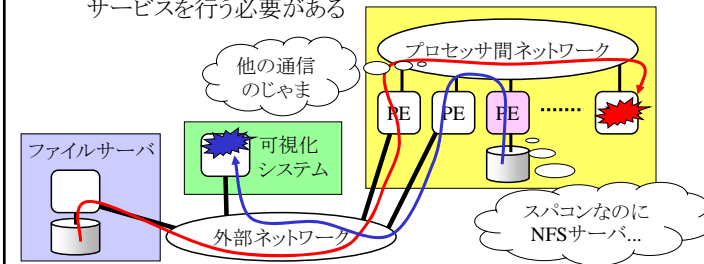
- HIPPI, Gigabit Ethernetなどでネットワークは高速になったものの...
  - ◆ディスクアクセスの負荷はすべてサーバに集中
  - ◆大容量のデータアクセスや多数のリクエストが発生するとネットワーク負荷が増大



29

### ネットワーク上のファイル共有の問題点(2)

- 並列型スーパーコンピュータなどでは...
  - ◆PEのうちの一部のみが外部ネットワークに接続
  - ◆外部のサーバにファイルを置くと、プロセッサ間ネットワークが他のジョブのデータで「汚れる」
  - ◆ファイルをローカルに置くと、可視化システムに対してNFSサービスを行う必要がある



30

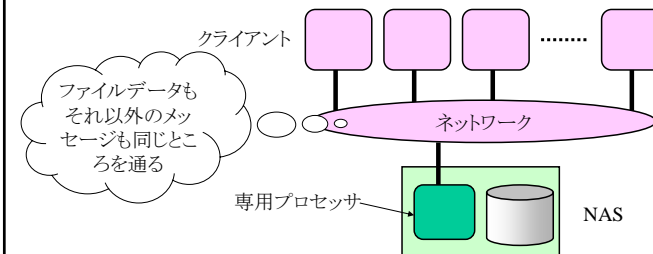
### 新技術を応用したファイル共有の形態

- NAS (Network Attached Storage)
  - ◆他のサービスも受け持つホストをファイルサーバにすべきでない
    - ファイルサーバをNFSサービスに専念させるのなら、いっそのこと専用機に
  - ◆計算機ネットワークにじか着けするディスク装置
  - ◆ファイル共有以外の通信サービスと経路を共有して使用
- SAN (Storage Area Network)
  - ◆大容量のデータ転送や多数のリクエストが殺到する場合のネットワーク負荷を軽減
  - ◆ファイル共有のためのネットワークを、それ以外の通信とは別の経路で
  - ◆最近注目されている技術
  - ◆FCの最初の応用か

31

### NAS

- 計算機ネットワークにじか着けするディスク装置
- NFS, CIFS, SMBといったファイル共有プロトコルを利用
- ただし、業界団体SNIA (Storage Networking Industry Association) が制定した用語の定義では、どんな種類のネットワークにつないでもかまわないことになっている



32



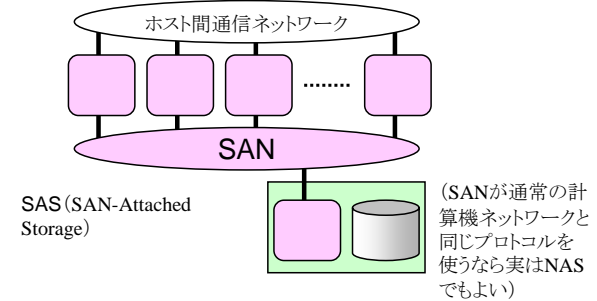
### NASの例

- この講義の第3回で紹介した以下の製品はNASである
  - ◆Auspex Netserver
  - ◆Netapp Multiprotocol Filer
- この他にも、多くのベンダーから多数の製品が発表されている

33

### SAN(1)

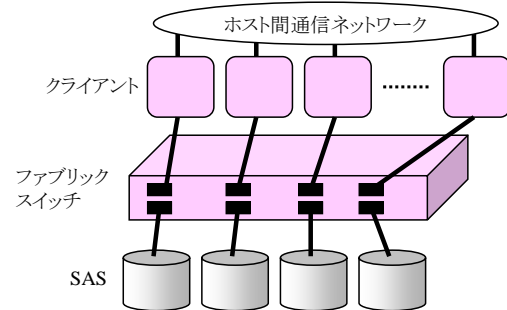
- 基本的な考え方
  - ◆ファイル共有のためのネットワークを、それ以外の通信とは別の経路で



34

### SAN(2)

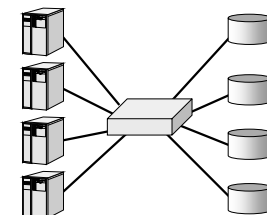
- FCファブリックなどを応用して、複数のディスクを共有することも可能
  - ◆ただし、共有ファイルへのアクセスの排他制御などはホスト側で処理する必要がある



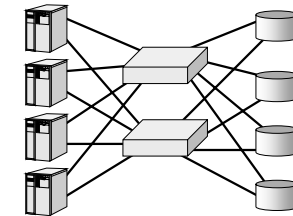
35

### SANのトポロジ

- スター型
  - ◆最も単純な構成
  - ◆スイッチの故障は、全システムに影響する
- デュアルスター型
  - ◆スイッチと接続を冗長化
  - ◆片方のスイッチが故障してもシステムは止まらない
  - ◆どちらのスイッチを使うかを切り替えることで負荷分散も可能
  - ◆スイッチがN台のときは N-wide スター型という



最近は8~16ポートのものをswitch, 32ポート以上のものをdirectorと呼び分けることもあるようだ



36

### SANのトポロジ(つづき)

- ツリー型
  - ◆スイッチをコアスイッチとエッジスイッチに分ける
  - ◆ノードはエッジスイッチにつながる
  - ◆コアスイッチの冗長化も可能
- リング型, メッシュ型
  - ◆多数のスイッチを, リング状または全対全で接続
  - ◆リングにコアを加えると, スター・リング (ring with star) 型

37

### SANのその他の効用

- ストレージの集中管理による管理効率の向上
- 複数のホストに分散するストレージを個別にバックアップ・リストアするよりも, 一括で行えるほうがよい
- SANバックアップ専用装置の利用
  - ◆バックアップ専用ホストやSANにじか着けられるテープライブラリ装置があれば, サービス用ホストやLANの負荷なしにバックアップ可能 (LAN-free backup)
- 異種ホスト間デバイス共有
  - ◆一つのディスク装置を, 異なるOSが動作する複数のホスト間で共有することが可能な製品も登場

38

### LAN-free バックアップ

- ホストやホスト間通信ネットワークに負荷をかけることなくバックアップ可能な構成

39

### 異種ホスト間ディスク共有

- SGI社のクラスタファイルシステム (CXFS)
- ADIC社のCentraVisionファイルシステム (CVFS)
- IBMソフトウェアのTivoli SANergyシステム

製品によってサポートしているOSに若干の違いはあるが...

40

## NAS vs. SAN?

- NASとSANを対立する技術として扱う意見
  - ◆NASのほうがSANよりも手軽
  - ◆SANのほうがNASよりもLANの負荷を軽減できる
- NASとSANは共存可能な関係にあるとする意見
  - ◆SANで結合されたストレージシステムも、外側(クライアント)から見ると、NASと同じ仕事をしてくれる
  - ◆同じセンターの中に両方あっても矛盾はしない
- NASとSANは将来統合されるのではないかとする意見
  - ◆FCなら同じ物理層の上にさまざまなプロトコルをのせることができる
  - ◆では、すべての通信をFCの上に
  - ◆FCスイッチを活用すれば、ホスト間の通信のバンド幅はデータ転送のじゃまを受けない

41

## まとめ

- ディスク利用形態の変遷
  - ◆集中処理, 分散処理, 協調処理
- 計算機とディスク(アレイ)の間の接続インタフェース
  - ◆SCSI
  - ◆HIPPI
  - ◆Fibre Channel
  - ◆GSN(Super HIPPI, HIPPI-6400)
- 接続トポロジ
  - ◆1対1, デイジーチェーン, ループ, スター
- NFS, CIFS, SMBによるファイル共有
- NAS
- SAN

42

## 次回予告

- 特別講義 12月19日 第2限
  - ◆「ストレージのテクノロジーと最新技術動向」  
ブロードコムコミュニケーションズ システムズ(株)  
小宮崇博氏
- 教室は変更になるかも知れません。
  - ◆この講義のwebサイト, および, 掲示に注意してください.

43