計算機システム概論
システム構成技術
2011/5/11

門林雄基

NAIST 奈良先端科学技術大学院大学

これまで学んだ事の関係性

2

プログラム	• アルゴリズム • データ構造
オペレーティングシステム	• CPUスケジューリング • メモリ資源の管理
計算機	• CPU • メモリ
デバイス	• 外部記憶装置 • 入出力装置

 11/05/11

講義のポイント

3

- オペレーティングシステムは最初どうやって入れるのか？
- プログラムが初めて動くまではどうなっているのか？
- 信頼性の高いシステムをどうやって構成するのか？

さまざまな種類のプログラム

4

- インストーラ
- アプリケーション
- ミドルウェア
- アプリケーションコンテナ
- サーバ
  - デーモン(UNIX)
  - サービス(Windows)
- クライアント
- ウォッチドッグ
- ブートローダ
- ファームウェア

今回の講義では定義しない

… これらを組み合わせてシステムを構成する。

システムが出来るまで

ハードウェアの構成

6

- 筐体の選定
  - 19インチラックマウント (1U, 2U, 4U etc.)
  - デスクトップ (E-ATX, micro-ATX, mini-ITX etc.)
  - カスタムシャーシ
  - 容積、電源容量
- 組み立て

- エアフローの考慮
- ケーブルの取り回し

 11/05/11

インストーラ(導入プログラム)が起動するまで

7

- 電源投入
- ファームウェア
- ブートローダ
- …
- カーネル
- デバイスの構成
- インストーラの起動

ファームウェア(またはBIOS)

8

- ファームウェア:基盤に備え付けの書き換え可能ROMに記録されており、電源投入時に動作するプログラム
- 自己診断シーケンス
  - 冷却系
  - メモリ不具合の有無
  - デバイス不具合の有無
- ブートデバイスの選択

ブートローダ

9

- ブートローダ：ファームウェアにより起動されるプログラムであり、オペレーティングシステムの起動（ブート）を単一目的とするもの
- OS未導入であるので、
  - 通常、CD, DVD等の可般型媒体から
  - ネットワークから導入する場合もある

 11/05/11

オペレーティングシステムの導入

10

- デバイスドライバの導入
  - ディスクの認識
  - ネットワークインターフェースの認識
  - その他デバイスの認識
    - キーボード、マウス、グラフィクス
- ディスクパーティションの構成
- ネットワークインターフェースの構成
- その他デバイスの構成
- ブートシーケンスの構成

 11/05/11

オペレーティングシステムの起動

11

- 電源(再)投入
- ファームウェア
- ブートローダー
- …
- カーネルの起動
- デバイスの構成
- サービスの起動

オペレーティングシステムの構成

12

- 通常、パッケージ群によって構成される
  - 必須パッケージ
  - その他のパッケージ
- パッケージの類型:
  - 特定のデバイスに対応するもの
  - 特定の構成方法に対応するもの(高信頼化など)
  - 特定用途に対応するもの

オペレーティングシステムの停止

13

（起動の逆）

- サービスの停止
- デバイスの停止
- カーネルの停止
- ファームウェアに戻る

より詳しくは

14

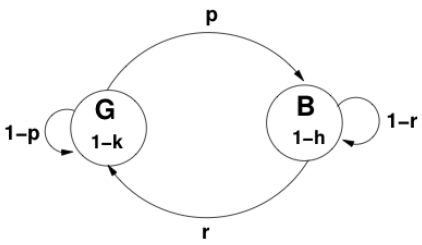
- 商用オペレーティングシステム、商用サーバ製品のマニュアルが詳しい
- 一通り手を動かしてみることを推奨する

システムの高信頼化

障害発生のモデル

16

- Bernoulli model: 確率p で障害発生
- Gilbert-Elliott model: 状態遷移によりモデル化



G: good state
B: bad state
1-k: error rate in good state
1-h: error rate in bad state
p, r: transition probability

Gilbert-Elliott model.
Source: G. Haßlinger and O. Hohlfeld, MMB 2008

 11/05/11

ごく単純な信頼性モデル: RAS

17

- 信頼性 (Reliability)
  - MTBF: Mean time between failure（平均故障間隔）
- 保守性 (Serviceability)
  - MTTR: Mean time to repair（平均修復時間）
- 可用性 (Availability)
  - A = MTBF / (MTBF + MTTR)
  - 年間ダウンタイムで表現することもある

- MTBF を長くするシステム構成技術とは？
- MTTR を短くするシステム構成技術とは？

高信頼化の手法（ハードウェア）

18

- 冗長化
  - 計算機まるごと
  - 電源
  - ストレージ
  - メモリ
  - CPU
  - ネットワーク
- ノイズの低減
  - 電源
  - EMI（電磁干渉）
  - ホコリ、静電気
- 排熱
  - エアフロー、液冷
  - 動作温度の監視

計算機の冗長構成

19

- duplex system
  - 本番系のコンピュータとは別に予備系のコンピュータを準備しておく形態
  - **ホットスタンバイ**:予備系のOSもプログラムも主記憶にロードされていて、すぐに切り替えることができる状態
  - **ウォームスタンバイ**: OSは立ち上がっているが、本番系に切り替えるには、処理に必要なプログラムを主記憶にロードする必要がある状態
  - **コールドスタンバイ**: 予備系に電源が入っていない状態、もしくは、現状の処理をすべて停止して、再度立ち上げが必要な状態
- dual system
  - 2系列(それ以上)のコンピュータシステムが常に同じ処理を実行し、お互いの結果を確認しながら処理を進める形態

電源の冗長化

20

- 冗長電源ユニット
- 電源レール複数化
  - 19インチラック

Source: Supermicro SuperBlade
3+1 redundant power supply modules

Source: 摂津金属工業 オンラインカタログ

ストレージの冗長化: RAID

21

- RAID: Redundant Array of Inexpensive Disks
- 大容量ストレージを安いディスクで実現する技術。
- 構成を工夫すれば、信頼性も提供できる。

- ディスクより大きなファイル、データベースを扱いたいときどうするか？
- concatenation
  - disk 1: block 1 ... block N
  - disk 2: block N+1 ... block 2N

RAID-0: striping

22

- (ディスク k 本を束ねたとして)
- disk 1: block 1, k+1, 2k+1, ...
- disk 2: block 2, k+2, 2k+2, ...
- disk k: block k, 2k, 3k, ...
- + ディスクアクセスの負荷分散が可能。
- + 大きなファイルの転送が、ディスク転送速度の総和で行える。
- - 信頼性が低い。ディスクが一つでも潰れたらデータ損失。

RAID-5: bitwise parity

23

- disk 1: block 1, k+1, 2k+1, ...
- disk 2: block 2, k+2, 2k+2, ...
- ...
- parity disk: parity(1..k), parity(k+1..2k), ...
- + ディスクがどれか一つ潰れても、他のディスクからデータ復旧が可能。たとえば:
  - disk 1: 1001
  - disk 2: 0101
  - disk 3: 1000
  - parity: 0100

RAID-5: bitwise parity

24

- - データ更新のさいに、データブロックとパリティブロックを更新しなければならない。パリティ更新中にクラッシュしたら間違ったデータを復旧してしまう危険性がある。
- → write-ahead logging

RAID-1: mirroring

25

- まったく同じ内容を複数のディスクにもつ。
- + 信頼性が最も高い
- - 速くならない、I/O バスの帯域を k倍占有する

メモリの高信頼化

26

- parity bit
  - 誤り検出
  - 訂正はできない

Data(3:0)	Odd Parity Bit	Even Parity Bit
0 1 1 0	1	0
0 0 0 0	1	0
1 1 1 1	1	0
1 1 0 1	0	1

Source: MIPS R4000 Microprocessor User Manual

- ECC: Error Correction Code
  - 誤り訂正
- IBM ChipKill, Sun Extended ECC 等
  - より強力な誤り訂正

⇒ 情報理論

CPU の冗長化

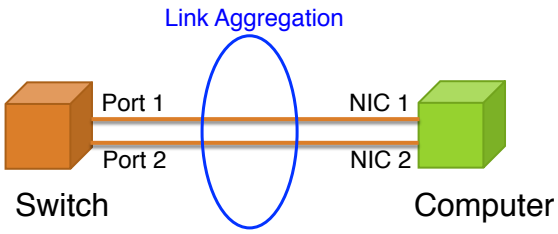
27

- メインフレーム
- フォールトトレラント・サーバ

ネットワークの冗長化

28

- IEEE 802.3ad (Link Aggregation)



*NIC: Network Interface Card

管理モジュールによる高信頼化

29

ILOM (integrated lights-out management)

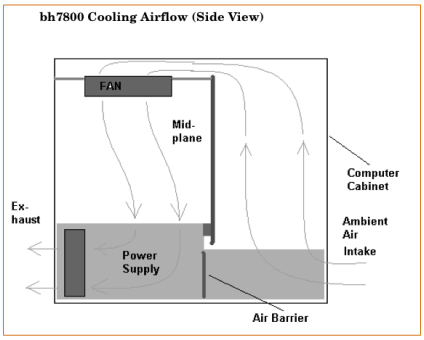
- 動作温度の監視
- 計算機の電源オフ／オン
- 管理ネットワーク経由での接続

- サーバ製品では必須の機能だが..

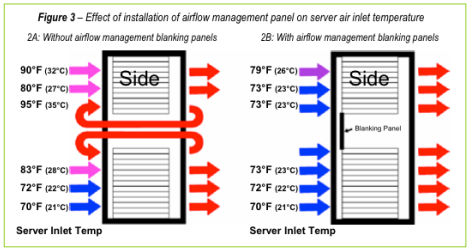
エアフロー

30

- ケース内エアフロー
- ラック内エアフロー



Source: HP blade server bh7800 installation guide



Source: APC white paper #44

液冷：熱伝達率に優れる

31

- スーパーコンピュータ
- 高密度クラスタ・システム

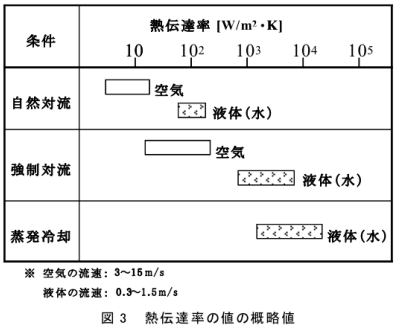
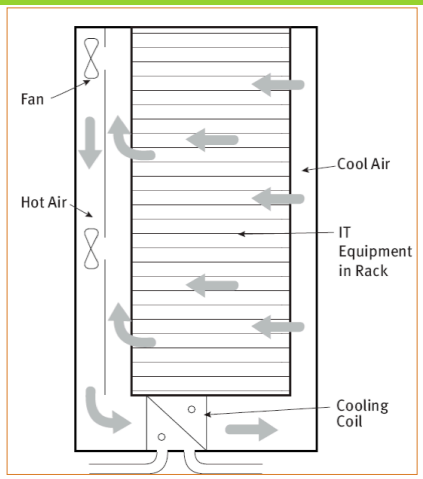


図３　熱伝達率の値の概略値

Source:
SANYO DENKI Technical Report No18 Nov. 2004



Source: Overview of Liquid Cooling Systems, LBL

 11/05/11

高信頼化の手法（ソフトウェア）

32

- ウォッチドッグ
- プロセスの冗長化
- データの冗長化

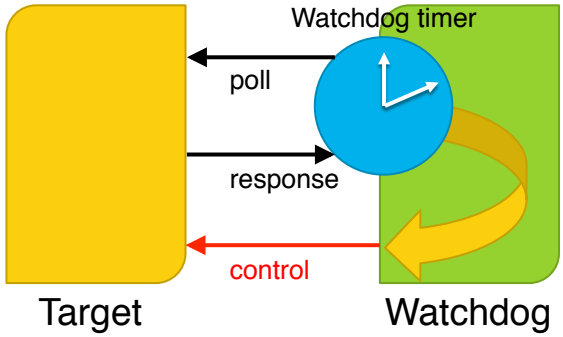
- バグの低減
  - テスト

 11/05/11

ウォッチドッグ

33

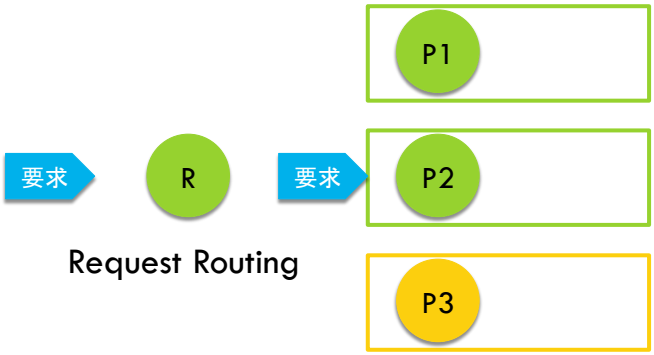
- 定期的な死活監視
- 一定時間、反応なければプロセス再起動



プロセスの冗長化

34

- 通常、ハードウェアの冗長化と組み合わせる
- 反応のあるプロセスに処理要求を転送



 11/05/11

データの冗長化

35

- 複製 (Replication)
- 冗長符号化 (Redundant coding)
  - 例：低密度パリティ検査符号 (LDPC)
  - ⇒ 情報理論

バグの低減

36

- Unit test: 関数などの小さな単位で期待通りに動くことをテスト
- Regression test: プログラム修正によるバグ発現の有無をテスト
- Fuzz test: 入力に異常データを混入させ、バグ発現の有無をテスト

バグが無ければ良いのか

37

- ソフトエラーの2大要因 (O'Gorman 1983)
  - α線 (alpha particle) - 材料による
  - 中性子線 (neutron) - 環境による
- ⇒ T. Karnik et al., "Characterization of Soft Errors Caused by Single Event Upsets in CMOS Processes", IEEE TDSC 1(2), 2004.
- ⇒ Radiation hardening

- 電磁干渉 (Electromagnetic interference)
  - トヨタ車の急加速事故でも当初疑われた（が、最終的に却下された）
  - 調査報告書: http://www.nhtsa.gov/UA

システム全体としての冗長構成

38

- ホットスタンバイ、コールドスタンバイ等は、ハードウェア・ソフトウェアいずれの障害に対応しているのか？
  - 製品によるため、要検討
- 主系に部分的な異常があった場合に代替系に切り替わらないこともある

- 完全無欠なシステムを作るのは難しい
  - → システム全体をテストする必要性
- Failover test
  - 切り替えができるか
- Fault injection test
  - 擬似的に障害を発生させテスト

まとめ：システム構成技術

39

- ハードウェアの構成
- ソフトウェアの導入

- システムの高信頼化
  - 障害発生のモデル
  - 信頼性の尺度： RAS (MTBF, MTTR)
  - MTBFを長くするシステム構成技術
  - MTTRを短くするシステム構成技術