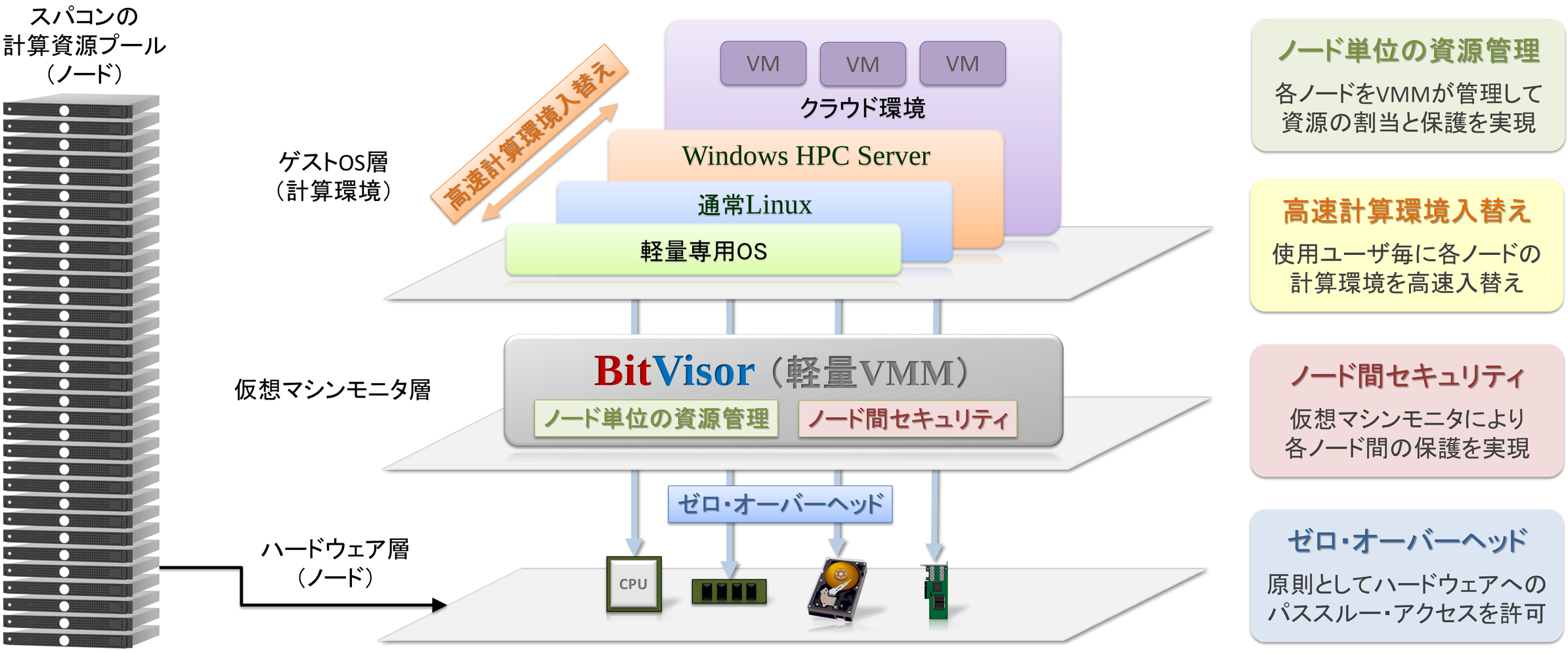
13-IS04

品川高廣 (東京大学)

次世代スーパーコンピュータ向けの軽量な仮想計算機環境の実現に向けた研究開発

BitVisor を活用したスパコン向け軽量仮想計算環境



ノード単位の資源管理
各ノードをVMMが管理して資源の割当と保護を実現

高速計算環境入替え
使用ユーザ毎に各ノードの計算環境を高速入替え

ノード間セキュリティ
仮想マシンモニタにより各ノード間の保護を実現

ゼロ・オーバーヘッド
原則としてハードウェアへのパススルー・アクセスを許可

近年のスパコンの構成

- 並列・分散型のコンピュータ
  - 多数のノードで構成される
    - 「京」は **88,128台**（102ノード×864ラック）
    - 「TSUBAME2」は **1,408台**（Thinノード，44ラック）
    - ⇒低遅延LAN接続の並列コンピュータ
  - 多数の拠点で構成される
    - 学際大規模情報基盤共同利用・共同研究拠点（JHPCN）
      - 7大学の計算機センターで構成
    - 革新的ハイパフォーマンス・コンピューティング・インフラ(HPCI)
      - 「京」及び国内38機関で構成
    - ⇒インターネット接続の分散コンピュータ


※HA8000クラスタシステム（T2K東大）HPより


※JHPCNパンフレットより

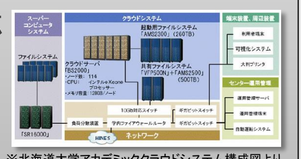
多数のノードの効率的な管理が必要

1

スパコンの用途の多様化

- 数値計算からクラウドまで
  - 数値計算の性質の多様性
    - ボトルネックが計算の種類によって異なる
      - CPU or メモリ or ディスク
      - ネットワーク帯域 v.s. 遅延
  - ビッグデータやホスティング対応
    - c.f. 北海道アカデミッククラウド
      - Hadoop, IaaS/Paas などをサポート

有限要素法や境界要素法などに基づく構造解析，気象予測，分子動力学，シミュレーション天文学，最適化問題，金融工学のような大規模数値解析に基づくシミュレーション，…
※Wikipediaより


※北海道大学アカデミッククラウドシステム構成図より

2

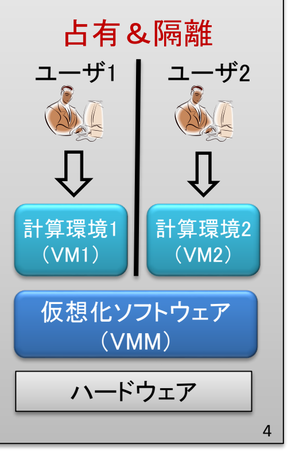
単一計算環境の限界

- 単一環境で全ての要求に答えることは難しい
  - OSに対する要求
    - 汎用OSの豊富な機能を使いたい（既存ライブラリ、Windows HPC Server）
    - 専用OSでオーバーヘッドを極限まで減らしたい（OSのジッタ等）
  - ストレージに対する要求
    - 手軽に使える大容量分散ファイルシステムを使いたい
    - ノードのメモリ間で高速に通信ができればよい
  - ネットワークに対する要求
    - 共有ネットワーク・インターネットにアクセスしたい
    - ハードウェアを直接叩いて超低レイテンシを実現したい

3

仮想化技術の活用

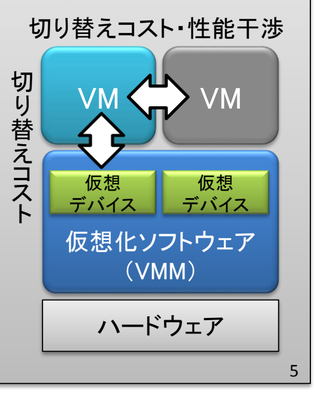
- 様々な計算環境を提供できる
  - 環境設定を自由に変えられる
    - ライブラリのバージョン指定
    - OS・システムのパラメータ設定
  - OS・カーネルを選択できる
    - 特定のバージョンの Linux カーネルの利用
    - Windows HPC Server の利用
    - 独自OS
- ユーザ間の保護も実現できる
  - VMごとに隔離された計算環境



4

従来の仮想化技術の問題点

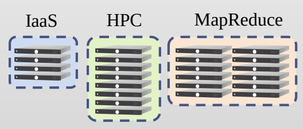
- 一定のオーバーヘッドが避けられない
  - VM間切り替えのコスト
    - コンテキストスイッチ
    - VM間スケジューリング
  - デバイス仮想化のコスト
    - VM-VMM間の切り替えが頻発
  - VMM自身のコスト
- 性能隔離が完全でない
  - 他のVMの性能が影響を受ける



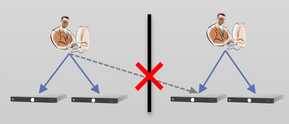
5

軽量VMMによるマシン管理

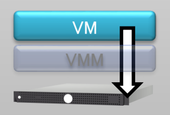
- ノード単位での資源管理
  - 「ノード群」をコンピュータとみなす
    - 比較的長期（時間・日・週）の時分割多重



- ノード単位でユーザに割り当て
  - 割り当てられたノードは専有利用可能
    - ノード間の隔離は軽量VMMで担保する



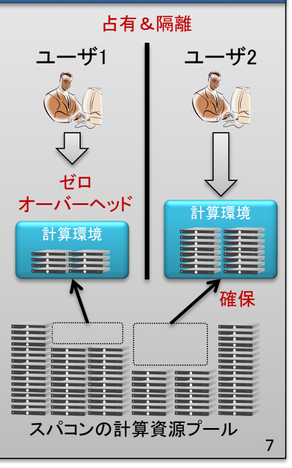
- 原則パススルーアクセス
  - VMMはハードウェアを仮想化しない
    - 最小限のアクセス制御のみ実施



6

本研究の最終目標

- 「スパコン as a Service」
  - 必要に応じた計算資源の確保
    - 数台から数百台まで on demand で利用
  - ノードの完全専有
    - ユーザごとに好きな計算環境を選択可能
  - ユーザ間の完全な隔離
    - セキュリティ＆製の上の影響を排除
  - オーバーヘッド≒ゼロ
    - 仮想化のコストを限りなく0に近づける



7

「BitVisor」の活用

- 純国産の軽量VMM
  - オープンソース
    - 修正BSDライセンスで公開
  - 小型軽量
    - コア部分のコードは約3万行
  - CPUの仮想化支援機構を活用
    - Intel VT / AMD-V対応
  - OS非依存・高い完成度
    - Windows, Linux等が動作
  - 研究プラットフォームとしての実績
    - HyperSafe, “Return-less” VMM, …

http://www.bitvisor.org/

8

BitVisor による高速計算環境入替

- 透過的ネットワークブート
  - ローカルブートを前提としたOSをネットワークブート
  - ATAへのアクセスを ATA over Ethernet(AoE) でサーバへ転送
- バックグラウンドインストール
  - ゲストOSの動作と並行してOSイメージを書き込み
  - デプロイまでのディレイを削減
- デバーチャリゼーション
  - インストール完了後，実行中にVMMの機能をOFFにする
  - 完全にゼロ・オーバーヘッドでの実行

9