DEOS-FY2009-PP-04

ディペンダブルサーバ構築要素技術
― DEOSで実現する簡単・安心・省エネサーバ ―

東京大学
石川裕

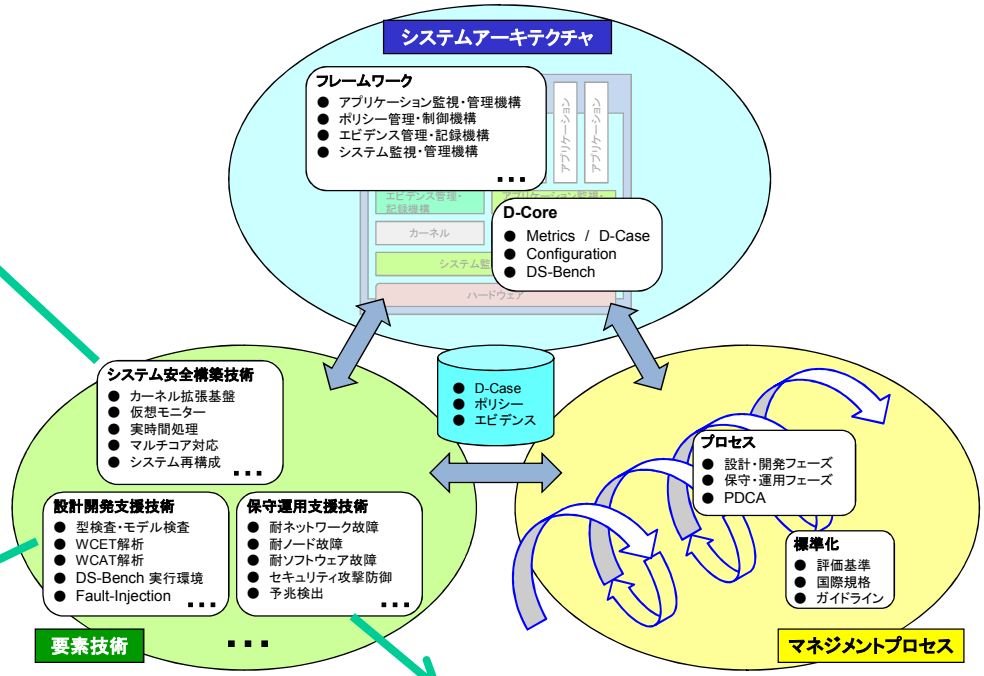
目次

- 石川チーム成果物の位置づけ
- 中間成果全体デモにおける位置づけ
- 研究の背景：サーバの利用形態
- シングルIPアドレス機構(SIAC)
  - 提供ディペンダビリティ
  - D-Core&他の要素技術との関連
  - 利点：拡張性、可用性、省電力性
  - 機構の概要
  - 既存システムとの比較
  - 今後の課題
- デモの紹介
- 重要な成果一覧
- まとめ

DEOS研究領域全体における石川チームの位置づけ

システム安全構築技術

開発モジュール	チーム名
仮想モニタ モニタリング(VMO) マルコア制御(VMC)	中島
P-Bus Core 論理分割(LPAR)	石川



設計開発支援技術

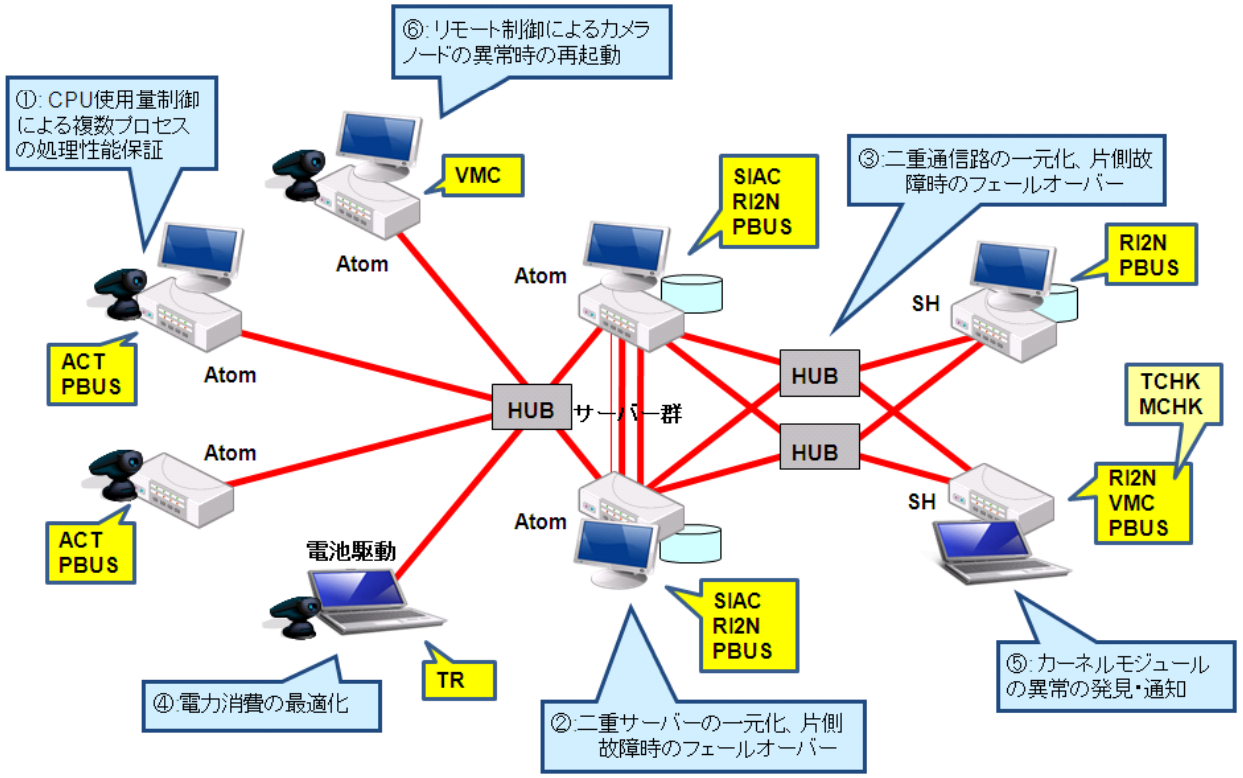
開発ツール	チーム名
型検査・モデル検査(TCHK/MCHK)	前田
最悪実行時間予測(RETAS)	石川
電力使用量予測(GREEN)	徳田
Fault Injection (D-Cloud)	佐藤
ディペンダブルシステムベンチマーク実行環境(DS-Bench)	石川

保守運用支援技術

開発モジュール	チーム名
動作時間予約機構(TR) 耐故障ネットワーク機構(SCTP+FHO)	徳田
耐故障ネットワーク機構(RI2N/PEACH)	佐藤
アカウンティング機構(ACT)	中島/センター
シングルIPアドレス機構(SIAC)	石川

中間成果全体デモにおける位置づけ

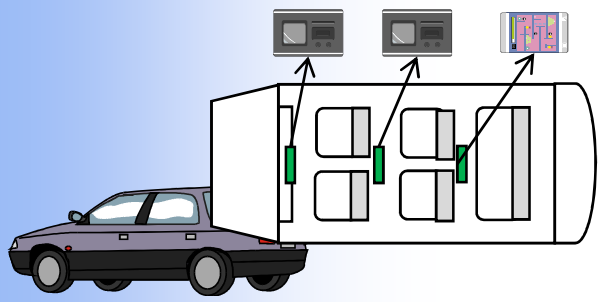
- P-Bus
  - 安全なOS機能拡張基盤であるP-Bus上で動作している耐故障ネットワーク機構であるRI2Nがデモで使われています
- シングルIPアドレス機構(SIAC)
  - シングルIPアドレス機構(SIAC)が提供する機能の一つであるサーバの可用性(Availability)をデモします



研究の背景：サーバの利用形態例

利用形態例

- オフィス利用
  - 顧客・財務・文書データ等の蓄積・処理
  - 監視システム映像蓄積
- 家庭利用
  - 映像、写真、ライフログ等格納共有
  - インターネット配信：家族、親戚、友達、不特定多数への配信
- 特殊条件下での利用
  - 車載端末の情報処理
  - ロボット制御における情報処理



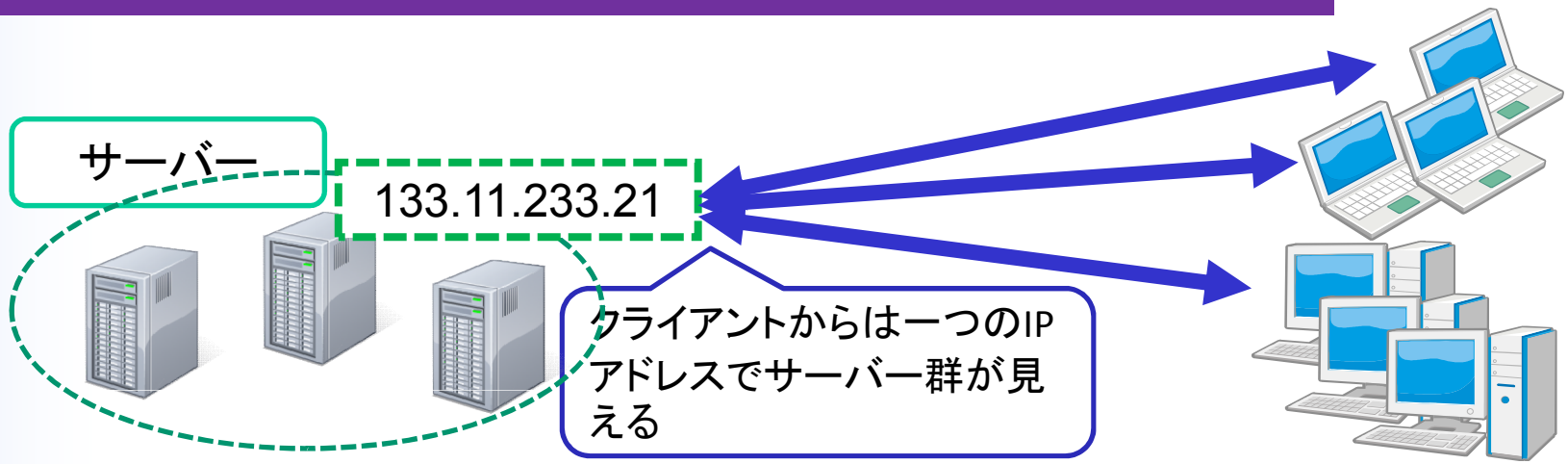
課題

- 処理能力と省電力の両立
  - 性能を求めると電力消費が増大
  - 省電力マシンは性能も低い
  - 負荷の増減
- 拡張性と可用性
  - ユーザ数増大に伴う性能向上＆可用性
  - 蓄積コンテンツ増加に伴う増強＆可用性
- オープンディペンダビリティの観点
  - 過負荷状態なのかエラー発生かの違い
  - 過負荷状態下による想定外動作

必要に応じて処理能力・容量を手軽に、
安心してエコに扱える要素技術の提供
シングルIPアドレス機構(SIAC)

シングルIPアドレス機構(SIAC)が提供するディペンダビリティ

サーバーノードを多数組み上げて1台のサーバに見せる技術。
クライアントからのリクエストは複数サーバノードの1つで処理される



- 拡張性
  - 必要に応じてサーバ台数を増やし、性能やディスク容量を増強可能
- 可用性
  - サーバを構成する要素の一部が故障しても全体のサービス継続
- 省電力性
  - 非力な省電力サーバの組み合わせによる性能保証と電力削減

- 既存技術との違い(概要)
  - 特別なハードウェアを必要としない
  - CPUを選ばない
    - 全体デモでは、x86系 CPU
    - 個別デモでは、PowerPC系 CPU
  - クライアントOSを選ばない

シングルIPアドレス機構(SIAC)とD-Core&他の要素技術との関連

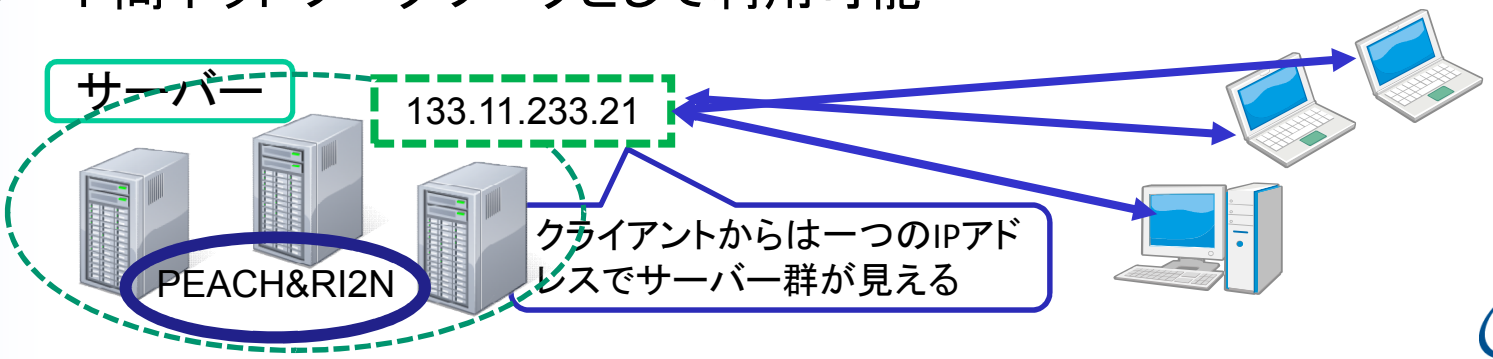
- D-Coreとの関連
  - D-Case要素技術の耐故障サーバとして利用
  - 耐故障サーバは、DS-Benchによる過負荷時挙動などで検査済
  - 新しい機器がシステムに接続される時、サーバは、接続機器のディペンダビリティ阻害要因の有無を確認。本確認のために、接続機器がD-Caseのエビデンスを持っているか調べる

    例1) 接続機器は、DOSアタックできない機構を持っている(DOS-noattack)

    例2）接続機器は、使用消費電力〇W以下である(PWR-consumption)
    - 電源容量が限られている環境において、サーバと接続機器が同一電源を共有している場合。例えば、プリントサーバ印刷時に必要な電源容量を確保するために使えるだろう

将来計画

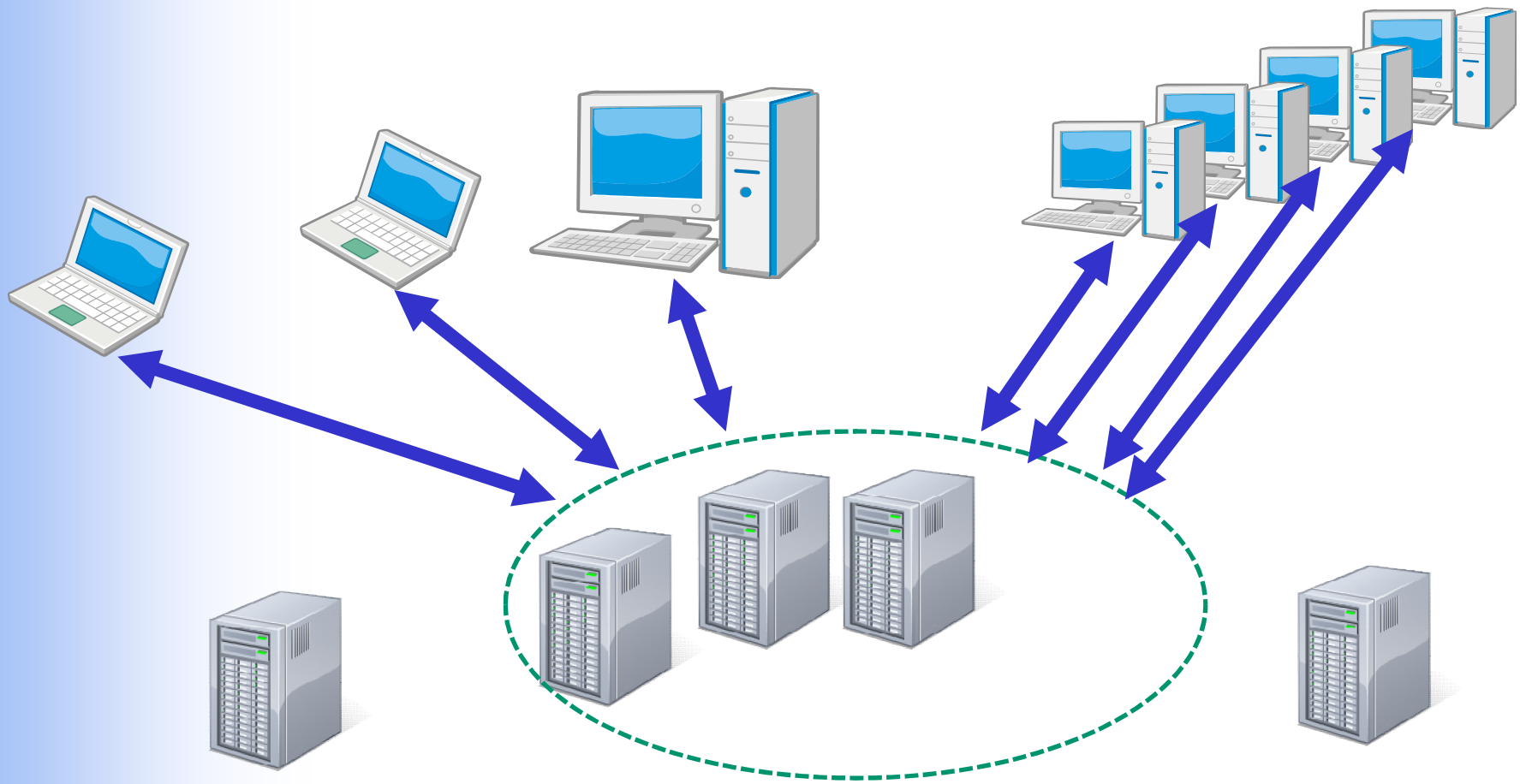
- PEACH&RI2N (佐藤チーム)との関連
  - サーバ内ノード間ネットワークワークとして利用可能



シングルIPアドレス機構(SIAC)の利点(1)

拡張性

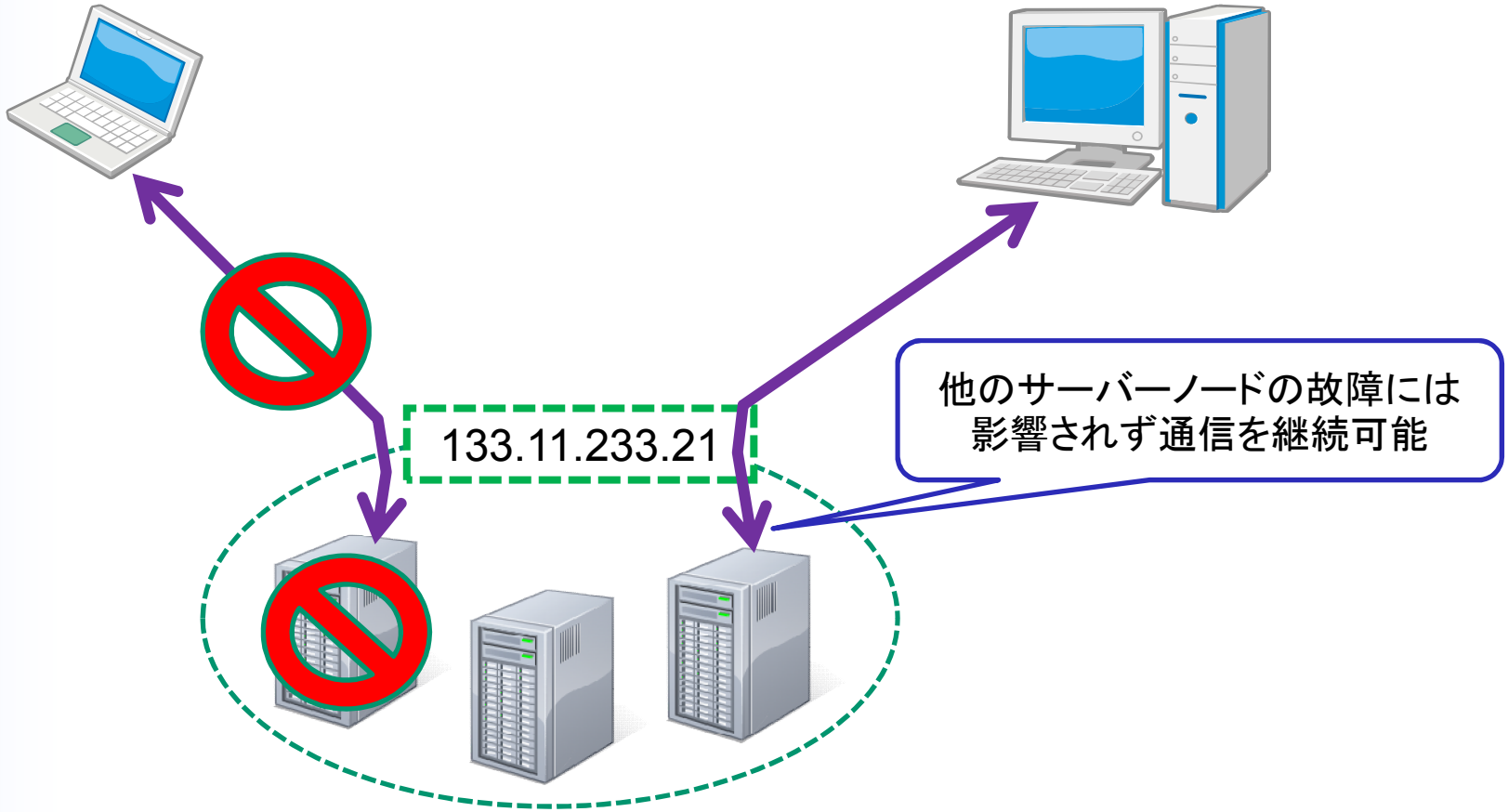
- クライアント側から透過的にサーバノード数を増やし総性能向上可能



シングルIPアドレス機構(SIAC)の利点(2)

可用性

- あるサーバノードが故障しても他のサーバノードは通信を継続可能

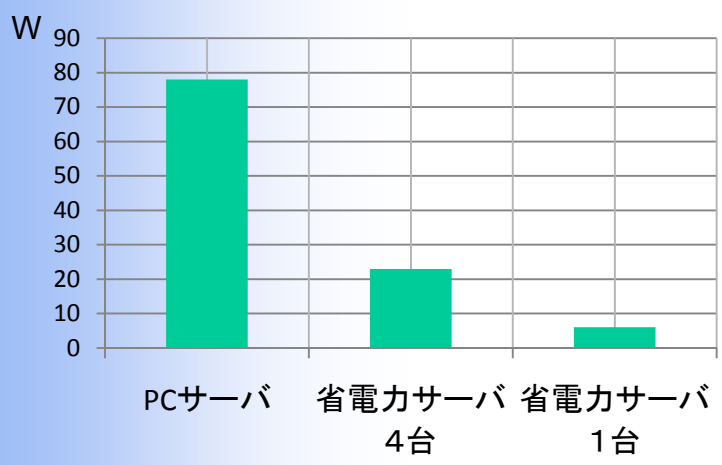


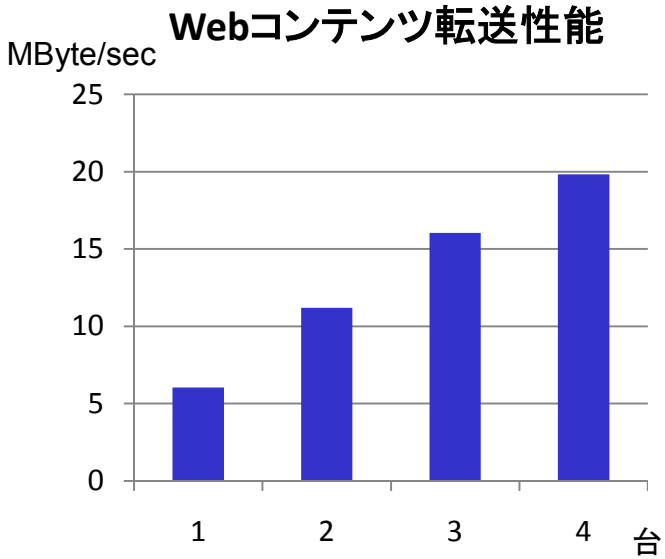
シングルIPアドレス機構(SIAC)の利点(3)

□ 省電力性

- 省電力サーバノードを複数用いることで、性能の総和を高めながら電力を削減することが可能

省電力サーバ: Plat'home OpenBlockS OBS266/128/16R
PowerPC 405GPr 266MHz, 128MB Memory, 32GB SSD
PCサーバ: DELL PowerEdge R410
Intel Xeon L5520 2.26GHz, 3GB Memory,
250GB 3.5inch SATA HDD





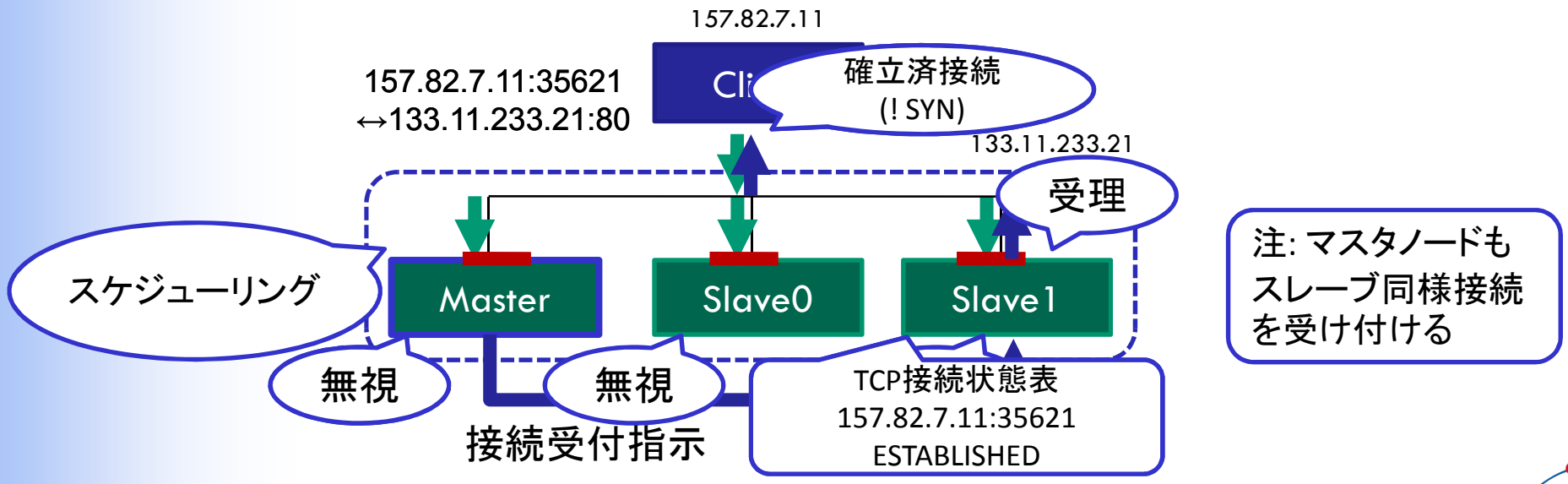
ポイント！

- 必要最低限の性能を有するサーバノードを選択
  例：地デジ映像なら16.8Mbps (2.1MB/sec)
- 将来処理量が増えたらノード追加で対応
- 耐故障機能も必要に応じて追加可能

従来、オーバースペックになりがちなサーバ仕様を的確なサーバ仕様にできる。
高負荷時の振る舞いは、DS-Benchによる検証を行い、安心・安全

シングルIPアドレス機構(SIAC)実現概要

- ブロードキャスト型シングルIPアドレス機構
- マスタノードの導入による新規TCPコネクションの割当管理
  - 任意のスケジューリングアルゴリズムを適用可能
- 接続確立後はマスタノードは関与せずに通信
  - マスタノードが故障しても既存コネクションは通信を継続可能
- マスタノードの耐故障性
  - マスターノード故障時は、他のノードがマスタノードとなる

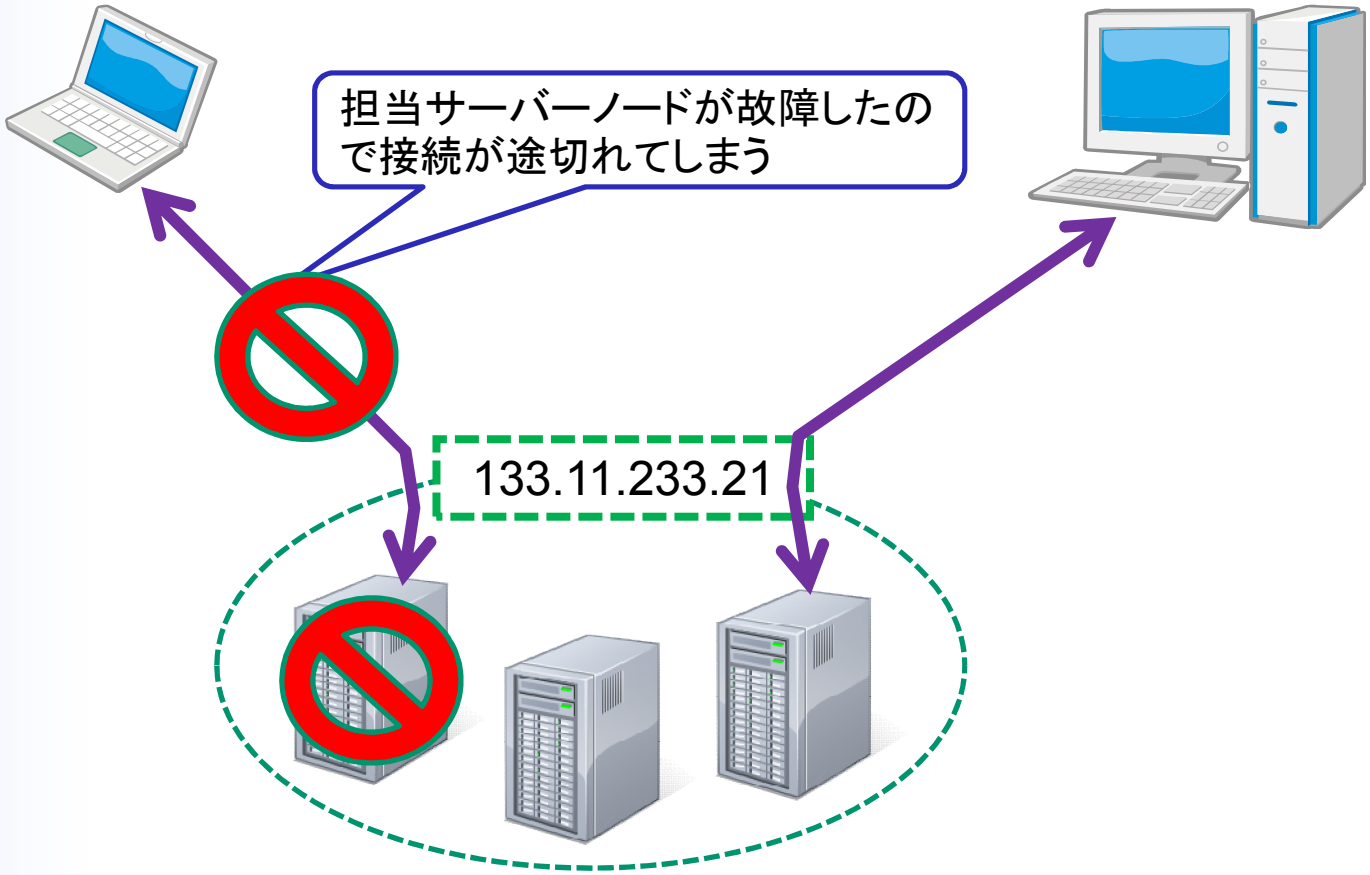


既存システムとの比較

- ロードバランサ＋サーバクラスタ
  - Webサーバ等の冗長化、高性能化に利用される
  - ロードバランサが故障すると利用できなくなる
- Hot-Standbyサーバ＋SAN(RAIDストレージ)
  - 高信頼サーバとしてよく用いられる形態
  - 高価で設置面積、消費電力ともに大
- クラスタ＋分散ファイルシステム
  - グリッドで使われているが、独自ファイルシステムプロトコル
  - Windows, MACなどクライアントごとに専用ドライバの開発およびインストールが必要
- 家庭用NAS
  - 省電力をうたう製品が増えている
  - 容量拡張性、性能拡張性に乏しい

今後の課題

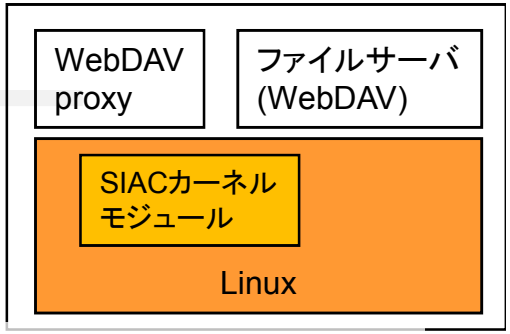
- 予兆なく故障したサーバノードに接続していたクライアントとの接続を継続する機構の開発

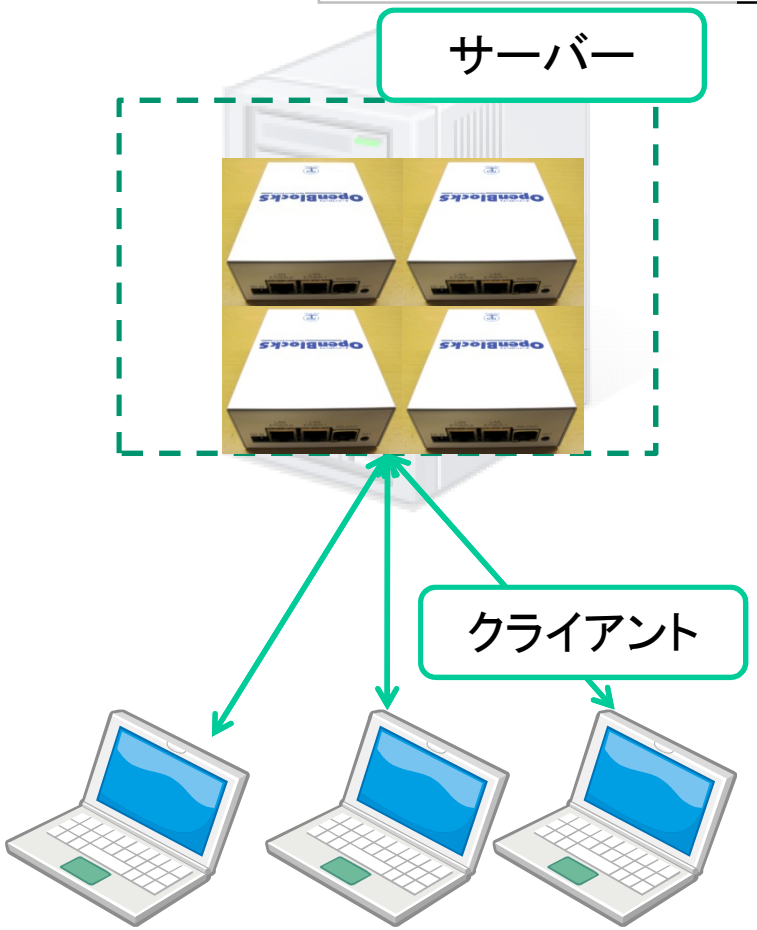


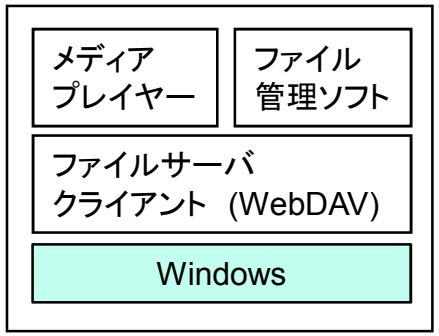
- Fail over時の性能チューニング

個別デモの紹介

- 省エネルギー分散ファイルサーバ
  - **Windows**から利用できるファイルサービスの一つである**WebDAV**プロトコルを実現
  - シングル**IP**アドレス機構によるクライアントからは1台の**WebDAV**サーバとして見える
  - ハードウェア
    - Plat'home OpenBlockS OBS266/128/16R
      PowerPC 405GPr 266MHz,　128MB Memory, 32GB SSD

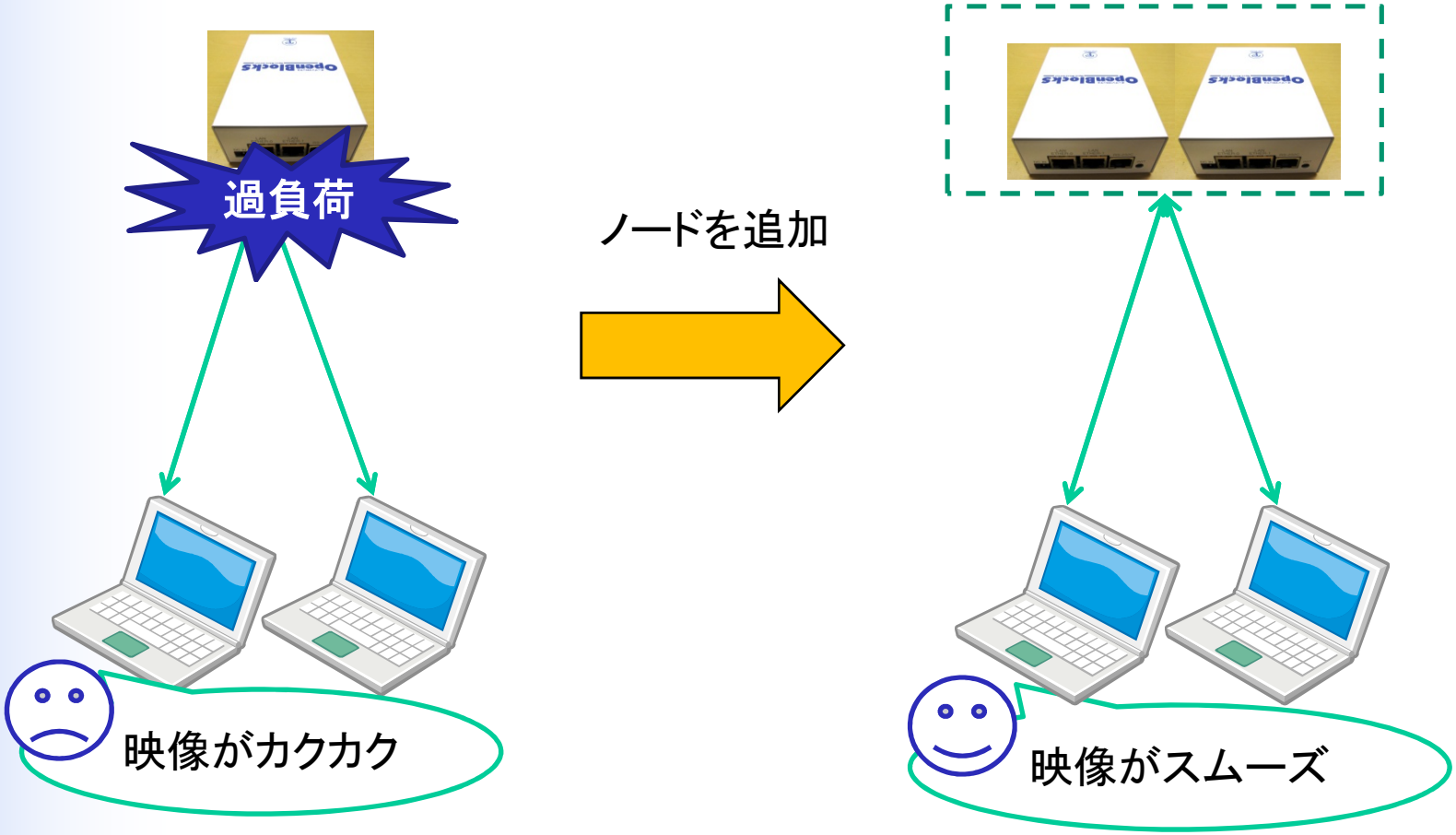






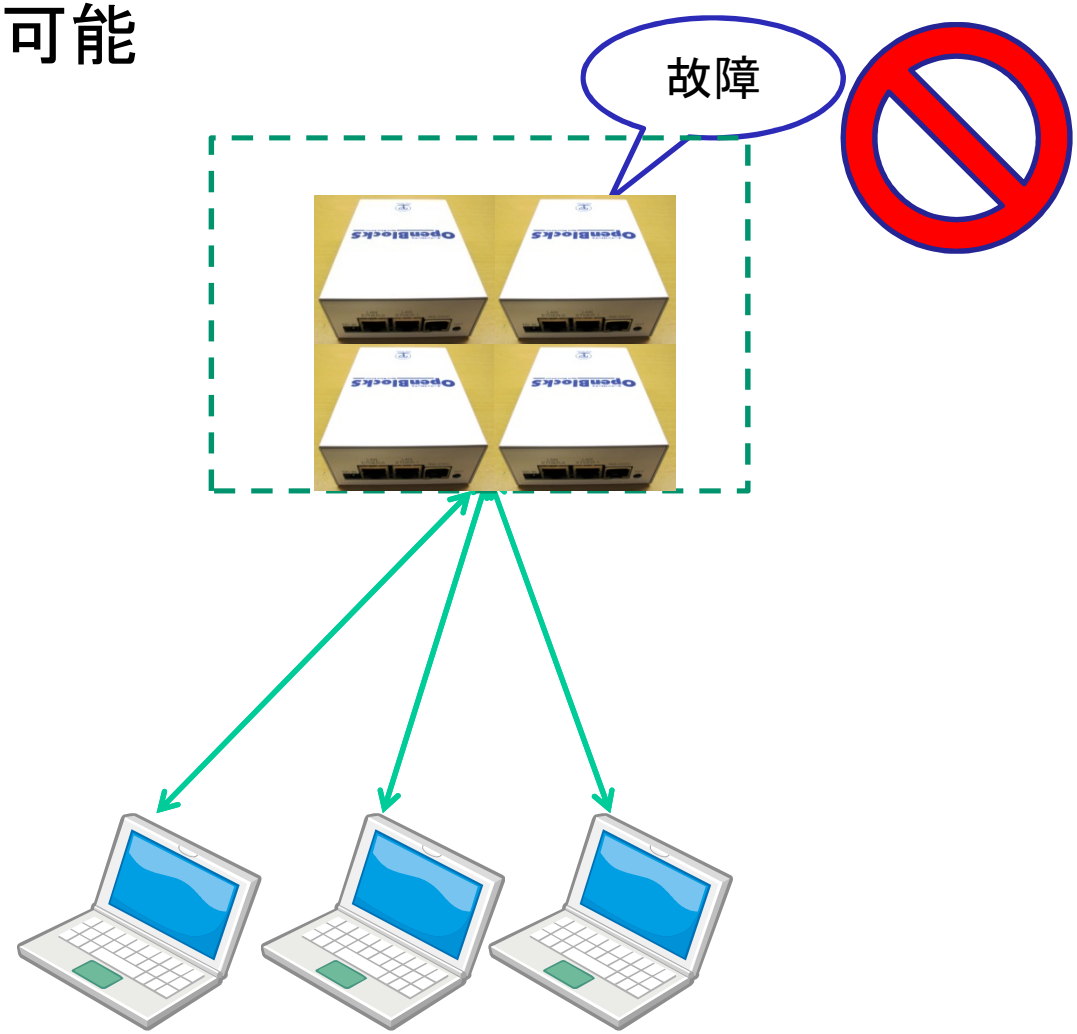
個別デモ：拡張性

- サーバノード数を増やすことによって合計性能を向上



個別デモ：耐故障性

- １台のサーバノードが故障しても
  データにアクセス可能



重要な成果

査読付き国際会議への採録

- Masato Sakai, Hiroya Matsuba and Yutaka Ishikawa ``Fault Detection System Activated by Failure Information,'' Proceedings of the 13th Pacific Rim International Symposium on Dependable Computing (PRDC'07), pp. 19 – 26, Melbourne, Australia, December 2007
- Hajime Fujita, Hiroya Matsuba, Yutaka Ishikawa, "TCP Connection Scheduler in Single IP Cluster", 8th IEEE International Symposium on Cluster Computing and the Grid (CCGRID'08), pp. 366-375, May 2008
- Taku Shimosawa, Hiroya Matsuba, Yutaka Ishikawa, "Logical Partitioning without Architectural Supports", 32nd IEEE International Computer Software and Applications Conference (COMPSAC 2008), pp. 355-364, 2008

受賞

- 藤田肇、平成20年度情報処理学会コンピュータサイエンス領域奨励賞
- 下沢拓、第109回情報処理学会OS研究会(平成20年8月) 最優秀学生発表賞

おわりに

- SIAC (Single IP Address Cluster)
  - 複数の計算機を単一のサーバに見せる技術
    - 省エネ
      - 従来、オーバースペックになりがちなサーバ仕様を的確なサーバ仕様にできる
    - 簡易性
      - ニーズに応じて構成が簡単に変えられ、拡張が可能
    - 安心
      - 耐ハードウェア故障を提供
      - 高負荷時の振る舞いは、DS-Benchによる検証を行い、安心・安全
      - クライアントに対するディペンダビリティ要件も動的に検査（将来）

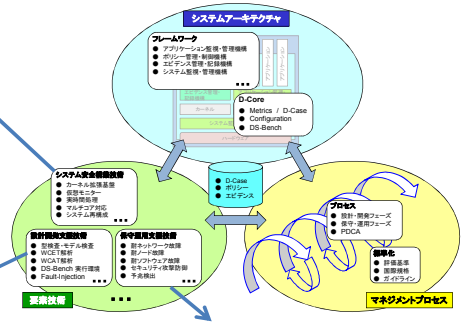
- 応用例
  - ファイルサーバ
  - ストリーミングサーバ
  - WEBサーバ

そのほかの石川チーム中間成果

- **RETAS最悪実行時間予測ツール**
- **DS-Benchディペンダブルシステムベンチマーク実行環境**
- **P-Bus Core**
- **論理分割 (LPAR)**

システム安全構築技術

開発モジュール	チーム名
仮想モニタ モニタリング(VMO) マルコア制御(VMC)	中島
P-Bus Core 論理分割(LPAR)	石川



設計開発支援技術

開発ツール	チーム名
型検査・モデル検査(TCHK/MCHK)	前田
最悪実行時間予測(RETAS)	石川
電力使用量予測(GREEN)	徳田
Fault Injection (D-Cloud)	佐藤
ディペンダブルシステムベンチマーク実行環境(DS-Bench)	石川

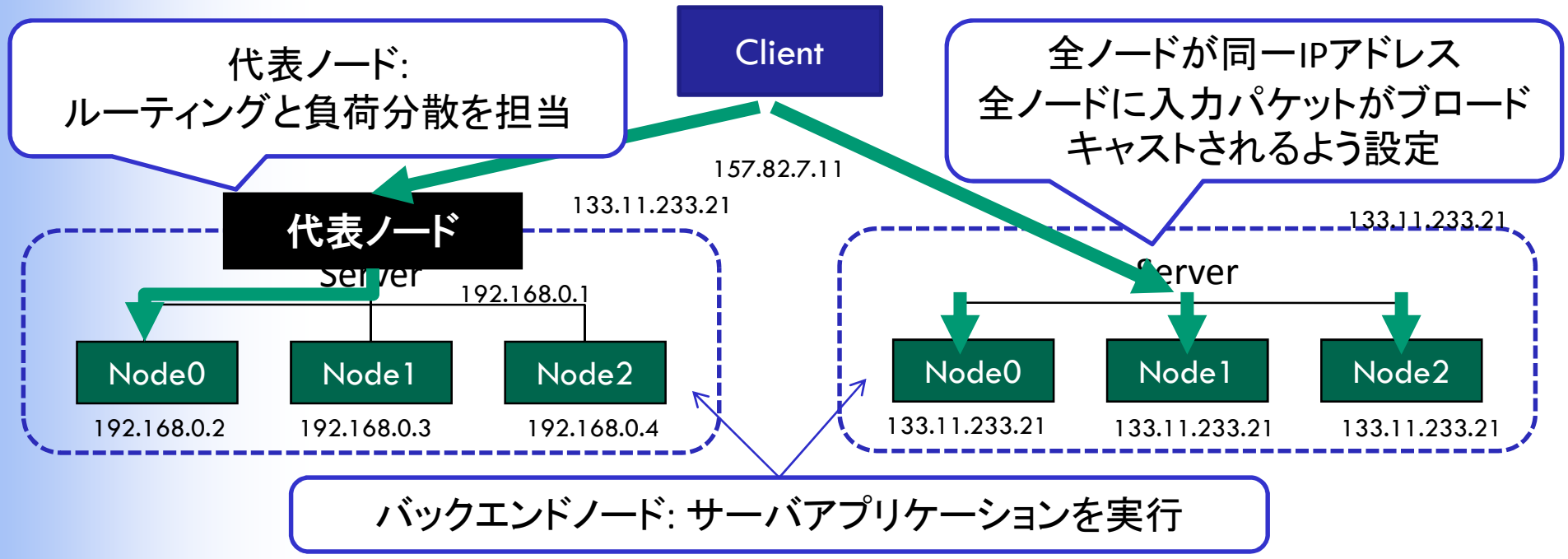
保守運用支援技術

開発モジュール	チーム名
動作時間予約機構(TR) 耐故障ネットワーク機構(SCTP+FHO)	徳田
耐故障ネットワーク機構(RI2N/PEACH)	佐藤
アカウンティング機構(ACT)	中島/センター
シングルIPアドレス機構(SIAC)	石川

APPENDIX

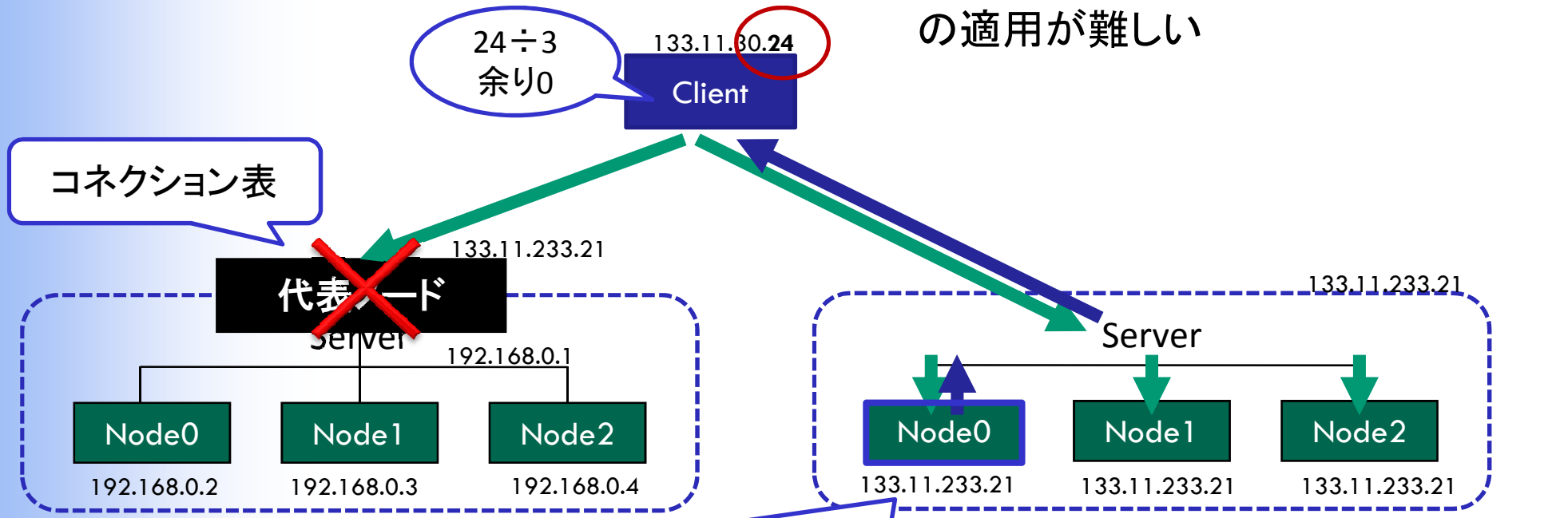
既存シングルIPクラスタ

- 代表ノード型
  - Linux Virtual Server[O'Rouke 01, Zhang 00]
  - SAPS[Mastuba 07]など
- ブロードキャスト型
  - Clone Cluster [Vaidya 01]
  - Windows NLB [Microsoft 02]など



既存手法の問題点

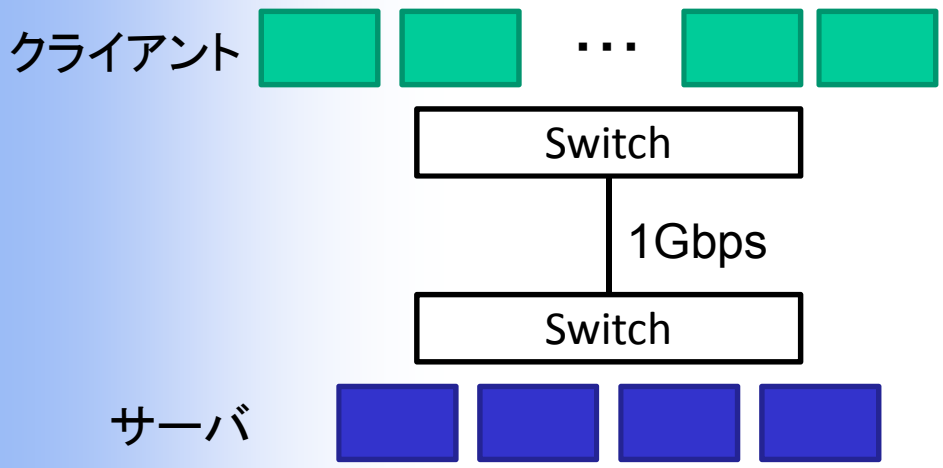
- 代表ノード型の問題
  - 代表ノードが単一障害点となる
- ブロードキャスト型の問題
  - 負荷分散が不完全[Vaidya 01]
    - パケット受信のためのルールが静的
    - 任意の負荷分散アルゴリズムの適用が難しい



静的ハッシュ関数の例: k ≡ i (mod N) k:キー i:ノード番号 N:ノード数
(クライアントのアドレス下位8ビットを全ノード数3で割り、余りがノード番号と一致したら受信)

評価概要

- 性能評価
  - 負荷分散に関する評価
- SPECweb2005 Supportベンチマークを使用
  - 4台のサーバ(Dual AMD Opteron 2.2GHz)に対し10台のクライアントからWebリクエストを発行
  - 同時webセッション 2,300
    - 2,300人の同時ユーザアクセス模擬



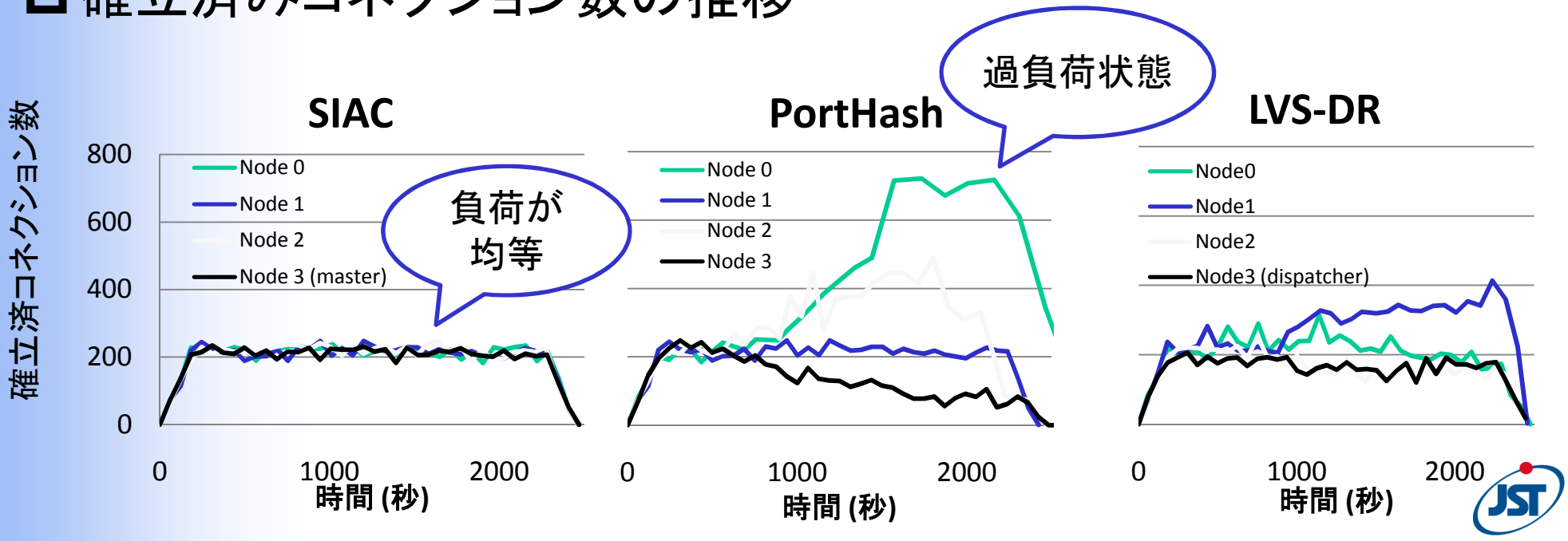
- SIAC+ least connection scheduling
  - 提案手法
- PortHash
  - 既存のブロードキャスト型クラスタと同様の方式
  - ポート番号をノード数で割った余りでコネクションを割当
- LVS-DR + least connection
  - Linux Virtual Serverによる代表ノード型クラスタ
  - バックエンドサーバからの返答は代表ノードを経由しない
  - 代表ノードもHTTPリクエストを処理
    - 計算資源の量を同一にするため

評価結果

◘ 総リクエスト処理数の比較 (30分間実行×3回の平均)

手法	概要	リクエスト処理数	割合
FTCS	提案手法	390,073	1
PortHash	ブロードキャスト型既存手法	344,488	0.883
LVS-DR	代表ノード型	387,639	0.994

◘ 確立済みコネクション数の推移



関連研究

- 静的ハッシュ関数に基づくブロードキャストクラスタ
  - **ONE-IP [Damani 97], Clone Cluster [Vaidya 01], Windows NLB [Microsoft 02]**
- 柔軟な負荷分散を目指すブロードキャストクラスタ
  - **Hive system [Takigahira 02]**
    - **IPパケットを受信するルールを記した表を定期的に更新**
    - 表の更新に要するコストが大きいのではないか
  - **[Baek 04]**
    - 到着したSYNセグメントの数を数え、各ノードが順番に接続受付することでRound Robinを実現
    - **SYNセグメントを一部ノードが取りこぼす可能性について言及されていない**

参考文献:

[Microsoft02] Network Load Balancing Technical Overview.

[Baek04] Seungmin Baek, Hwakyung Rim, and Sungchun Kim. Socket-based RR scheduling scheme for tightly coupled clusters providing single-name images. Journal of Systems Architecture, 50(6):299–308, 2004.

[Damani07] O. P. Damani, P. E. Chung, Y. Huang, C. Kintala, and Y.-M. Wang. ONE-IP: techniques for hosting a service on a cluster of machines. In *Selected papers from the sixth international conference on World Wide Web*, pages 1019–1027, Essex, UK, 1997. Elsevier Science Publishers, Ltd.

[Matsub07] H. Matsuba and Y. Ishikawa. Single IP address cluster for internet servers. In *Proceedings of 21st IEEE International Parallel and Distributed Processing Symposium (IPDPS2007)*, 2007.

[Rourke01] P. O'Rourke and M. Keefe. Performance Evaluation of Linux Virtual Server. *LISA 2001 15th Systems Administration Conference*, 2001.

[Takigahira02] T. Takigahira. Hive server: high reliable cluster web server based on request multicasting. In *Proceedings of The Third International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'02)*, pages 289–294, 2002.

[Vadiya01] S. Vaidya and K. J. Christensen. A single system image server cluster using duplicated MAC and IP addresses. In *Proceedings of the 26th Annual IEEE Conference on Local Computer Networks*, pages 206–214, 2001.

[Zhang00] W. Zhang. Linux Virtual Servers for Scalable Network Services. *Linux Symposium*, 2000.