

非対称ネットワークを隠蔽する 高速通信インフラストラクチャの 設計と実装

濱野 智行[†], 中田 秀基^{††, †}, 松岡 聡^{†, †††}

[†]: 東京工業大学, ^{††}: 産業技術総合研究所

^{†††}: 国立情報学研究所



アジェンダ

■背景・目的と要件定義

- 既存の隠蔽手法
- 新手法の提案と設計
- プロトタイプ実装
- 評価・考察
- まとめ

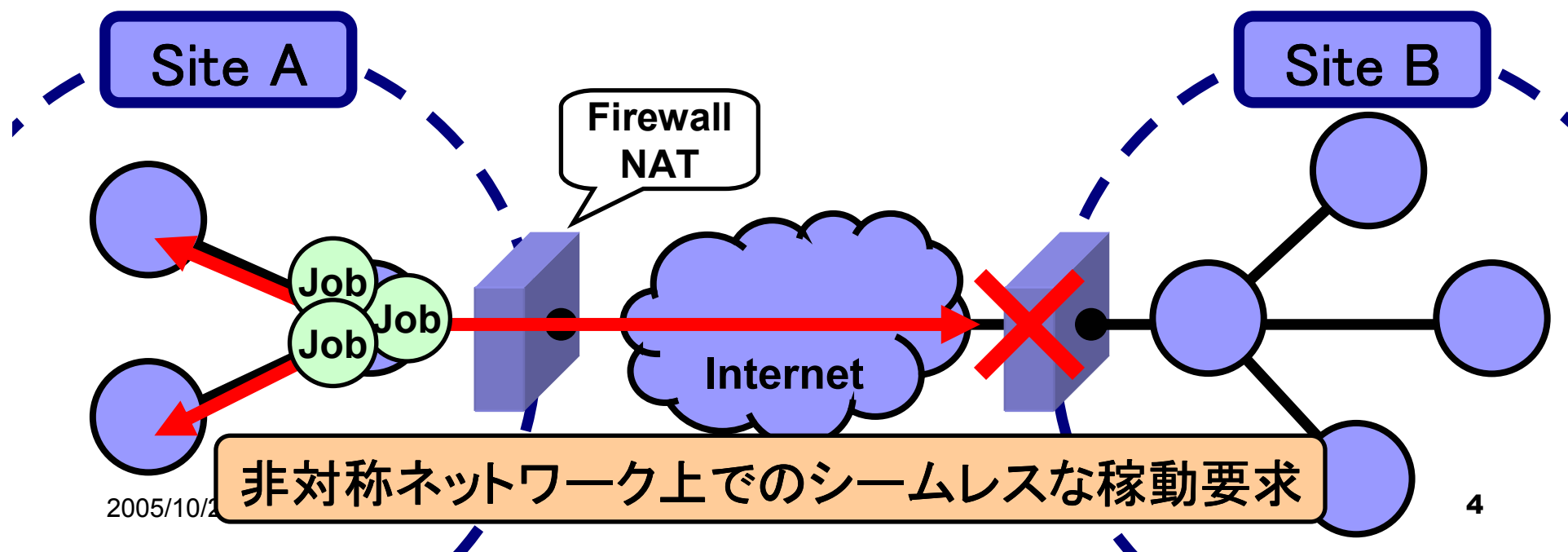


背景

- グリッド環境での広域分散計算が現実的に
 - 各サイト間での協調の必要性
- 非対称ネットワークがサイト間通信を妨害
 - ファイアウォール
 - (広義の)NAT
- 既存の非対称性を扱う研究
 - 同時にグリッドの要件を満たす必要性

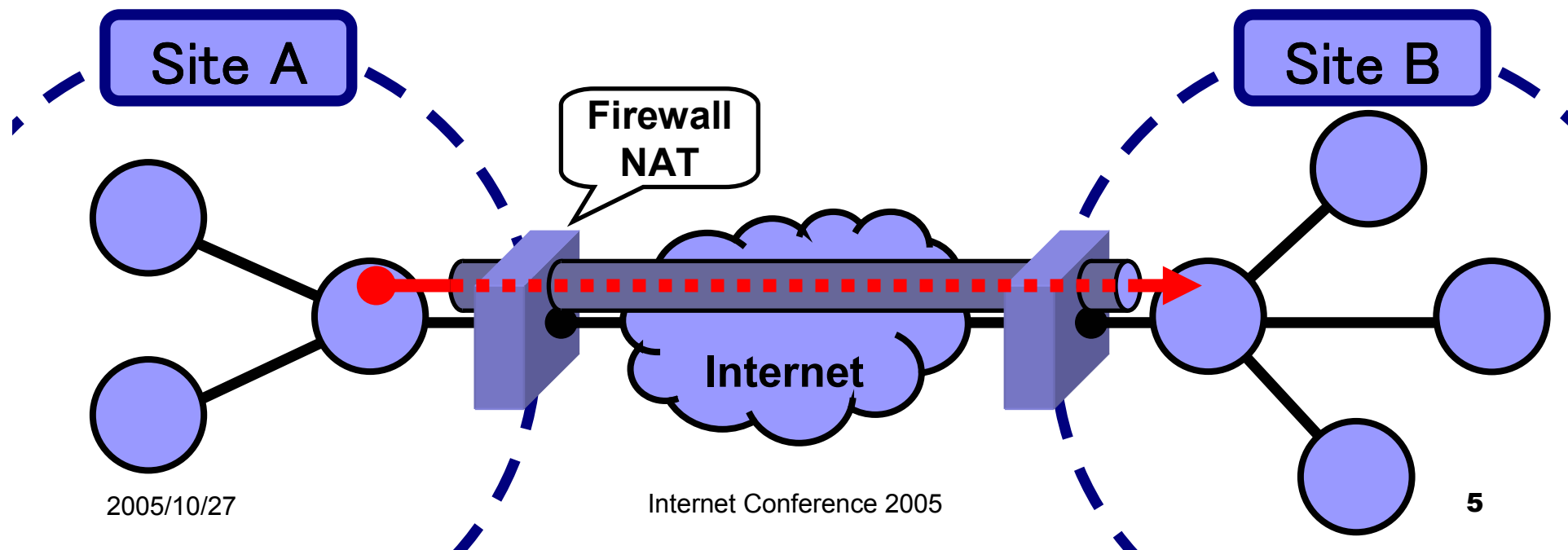
非対称ネットワークが問題になる典型例

- Condor [Livny et al. '88], Jay [Machida et al. '04] などのジョブスケジューリングシステム
 - 異なるプライベート空間に配置されたホスト間の通信妨害



目的


- 非対称ネットワークを意識させない高速通信インフラストラクチャの構築
 - グリッドのインフラとしての要件を充足
 - 対象はファイアウォールとNAT





グリッドの通信インフラとしての要件

- セキュリティ
 - 証明局から認められたユーザ・リソースのみ参加可能にするための認証・認可機能
 - 通信の傍受を防ぐ暗号化機能
- サイトポリシー非依存
 - 様々なプラットフォームが存在するグリッド環境に対応
 - 権限に依らず動作
- 高通信性能
 - 他分野の実験・観測データの肥大化への対応要請
 - 高エネルギー物理学: LHC (Large Hadron Collider) プロジェクト
 - 天文学: 仮想天文台



アジェンダ

- 背景・目的と要件定義
- **既存の隠蔽手法**
- 新手法の提案と設計
- プロトタイプ実装
- 評価・考察
- まとめ



非対称性を生じない技術 - IPv6

■ 利点

- アドレス空間拡大によりアドレス枯渇対策用NATを排除可能

■ 欠点

- 導入可否はサイトポリシーに大きく異存
 - NATを内部トポロジ隠蔽に使用するポリシーの存在
- 設定コスト大
 - OSやルータのIPv6化が必要
 - IPv4との共存を考慮する必要性
- ファイアウォールに別途考慮の必要性

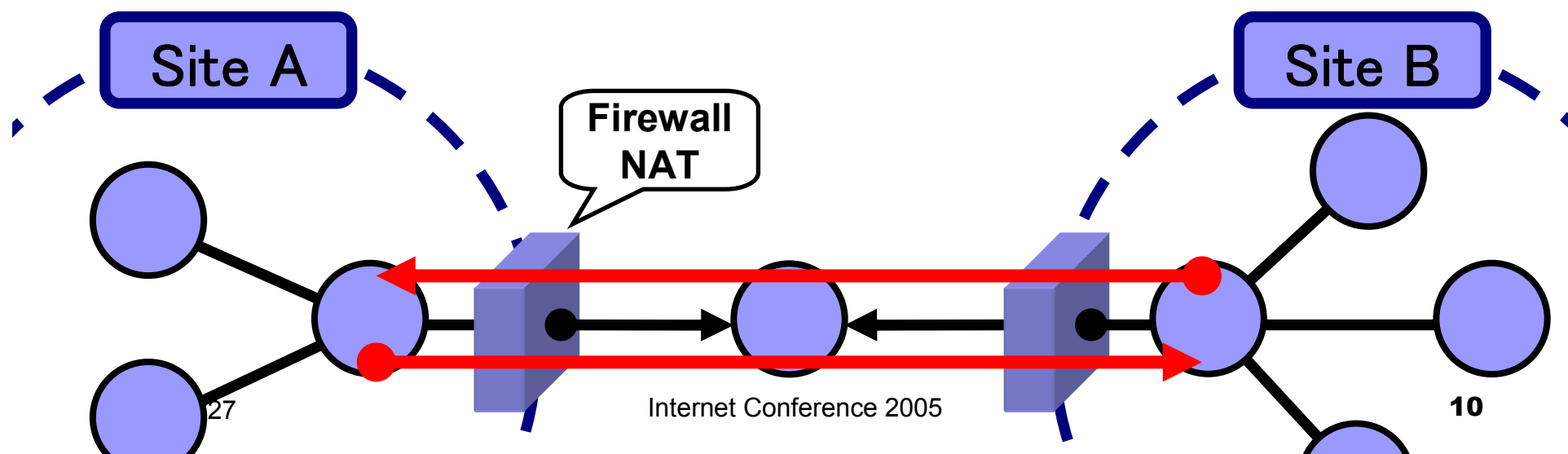


非対称性を隠蔽する技術

- 中間ホストによる通信のリレー
 - SOCKS [Leech et al. '96], GCB [Son et al. '03]
- NATフォワーディングルールの動的変更
 - DPF [Son et al. '03], RSIP [Borella et al. '00]
- UDP Hole Punching
 - TURN [Rosenberg et al. '03]

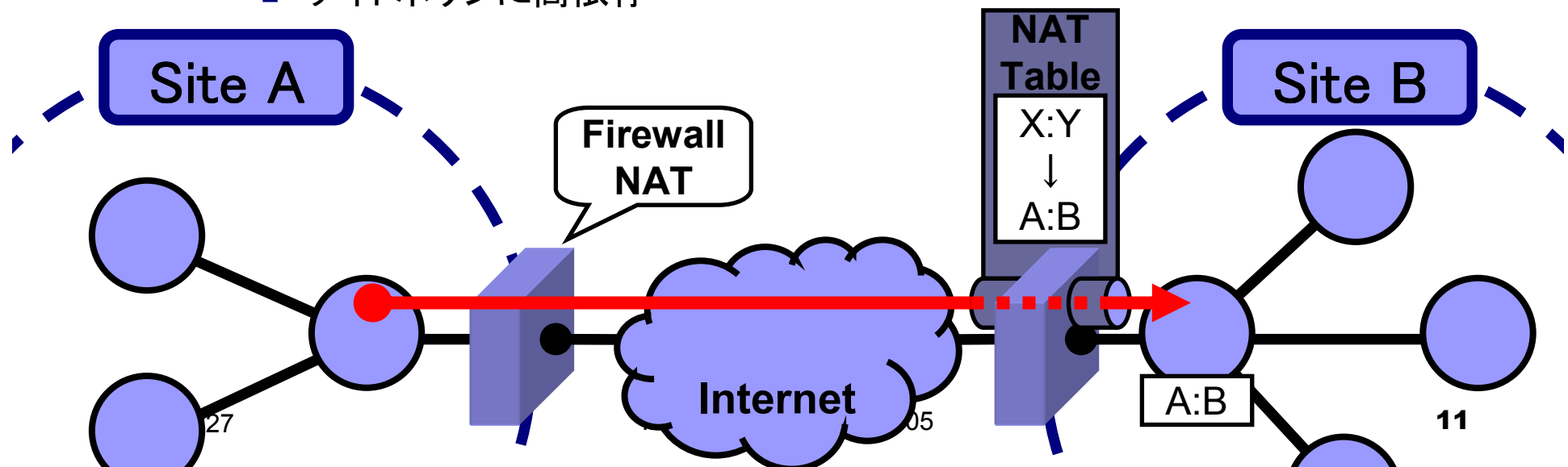
中間ホストによる通信のリレー

- 各サイトから接続可能な中間ホストが通信をリレー
 - 利点
 - サイトポリシーの制約を受けにくい
 - 欠点
 - リレーによる通信性能低下
 - UDPに別途考慮の必要性



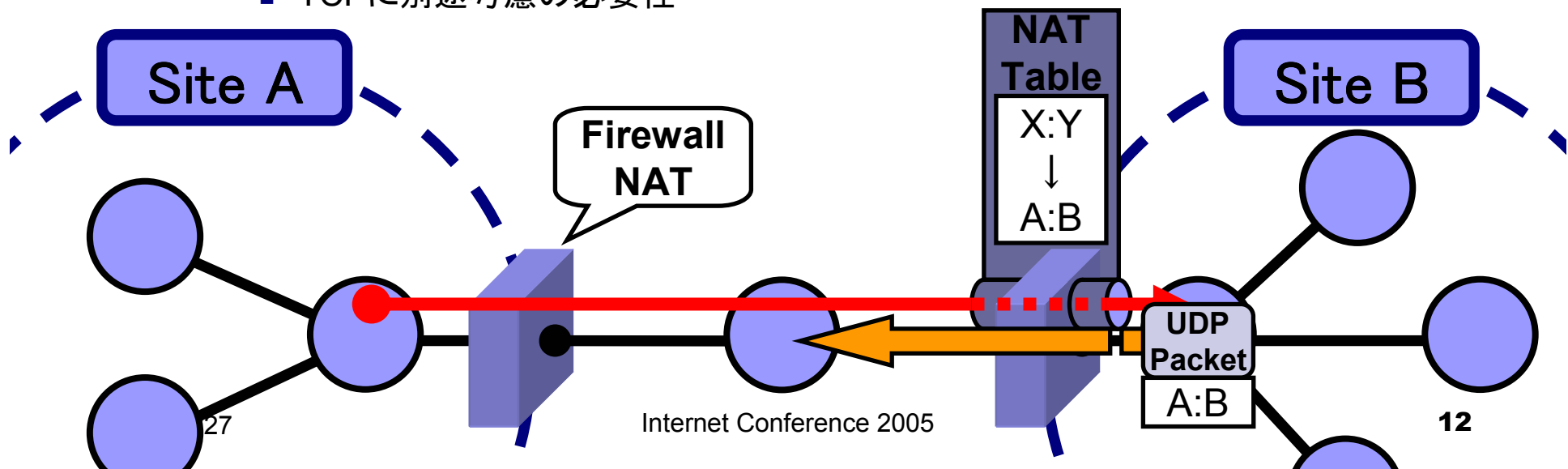
NATフォワーディングルールの動的変更

- セッション確立時、内部ホストに外部アドレスを対応付け、NATルールを動的に変更
 - 利点
 - 通信性能低下が小さい
 - 欠点
 - サイトポリシーに高依存



UDP Hole Punching

- UDPパケットを定期的を送信し、NATルールを維持
 - 利点
 - サイトポリシーの制約を受けにくい
 - 欠点
 - Symmetric NATでは使用不可
 - 中間リレーノードの存在により、通信性能低下
 - TCPに別途考慮の必要性



既存の隠蔽手法の比較

	接続性	ポリシー非依存	通信性能
中間ホストによるリレー	○	○	△
NATテーブル動的変更	○	×	○
UDP Hole Punching	△	○	△

NATテーブルを
変更可能な権限
が必要

リレーコストにより
通信性能低下

使用できない
NATが存在する

リレーコストにより
通信性能低下

これら3項目すべてを満たす手法は存在しない



アジェンダ

- 背景・目的と要件定義
- 既存の隠蔽手法
- **新手法の提案と設計**
- プロトタイプ実装
- 評価・考察
- まとめ

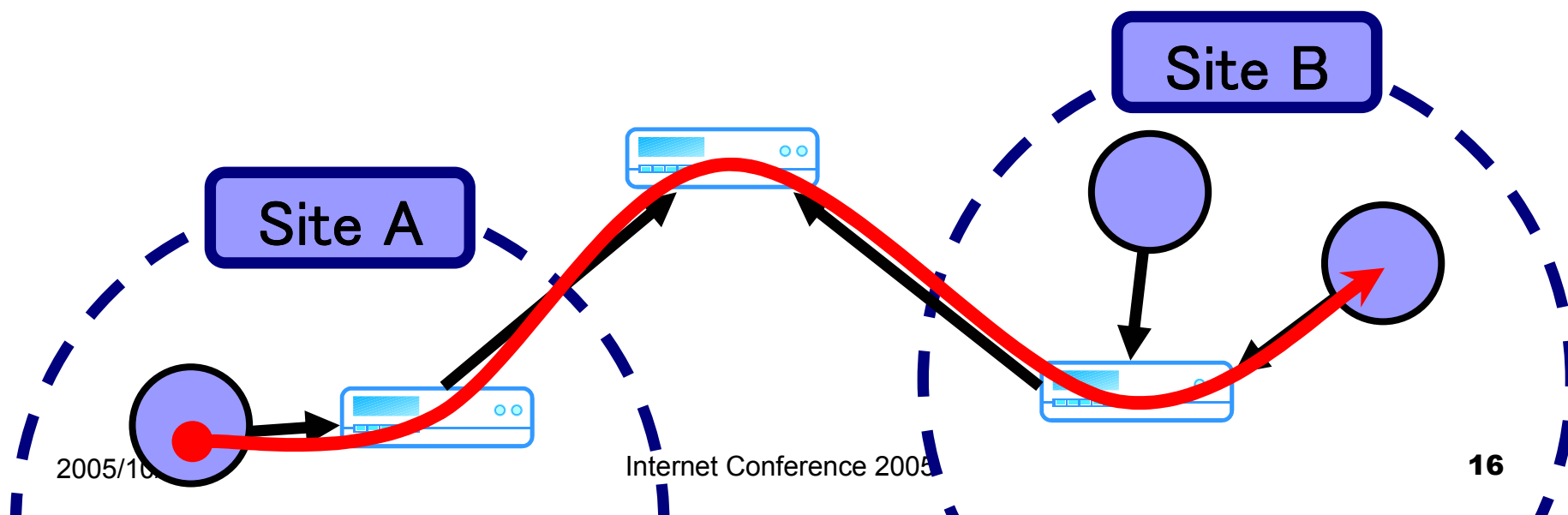
グリッド環境に適した非対称ネットワークを隠蔽する高速通信インフラストラクチャの提案

- 「中間ホストによるリレー」を適用
 - 接続性・サイトポリシー非依存を達成
 - 高通信性能は別の手段で達成
- セキュリティ機構の導入

	接続性	ポリシー非依存	通信性能
中間ホストによるリレー	○	○	△
NATテーブル動的変更	○	×	○
UDP Hole Punching	△	○	△
提案システム	○	○	○

システムの概要

- データをリレーするソフトウェアルータを配置し、オーバレイネットワークを構成
 - 通信性能向上はオーバレイネットワークの特徴を利用して達成する方針





本システムの課題

- オーバレイネットワーク構築
 - ネーミング: IPに依らない名前規則
 - ルーティング: 通信リレーの経路計画
- グリッドの要件
 - セキュリティ
 - サイトポリシ非依存
 - 高通信性能



設計(1/2) - オーバレイネットワーク構築

■ ネーミング

- 各ルータ・ノードが任意に決定
- 隣接ルータが接続時に名前の一意性を保証
 - 名前が既に存在する場合、接続を拒否

■ ルーティング

- 各ルータが保持する経路情報を定期的に隣接ルータに通知
 - 受信した経路情報をマージ
- トポロジ全体を把握した上で経路策定



設計(2/2) – グリッドの要件

- セキュリティ
 - PKI (Public Key Infrastructure)を使用
 - ホスト/ユーザの認証・認可
 - SSLにより通信を暗号化
- サイトポリシ非依存
 - Pure Javaで実装
 - 通常の権限で動作可能なシステム
- 高通信性能
 - ネットワークトポロジ全体を把握した上での経路選択
 - 最短ホップ数で到達可能

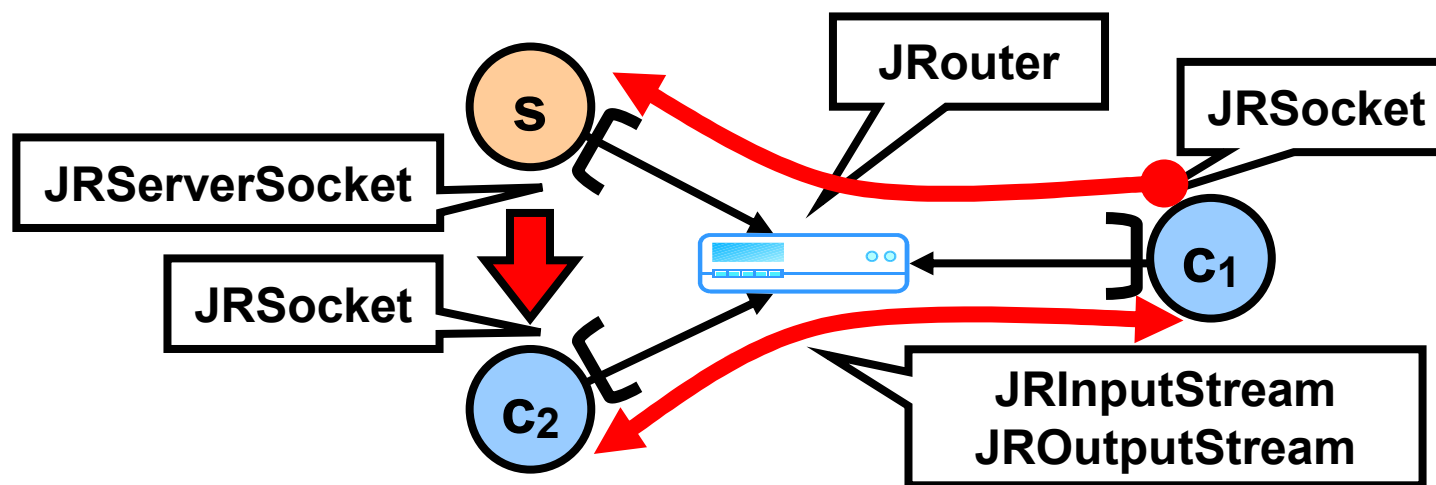



アジェンダ

- 背景・目的と要件定義
- 既存の隠蔽手法
- 新手法の提案と設計
- **プロトタイプ実装**
- 評価・考察
- まとめ

プロトタイプ実装


- 提案システムのプロトタイプJRouterを実装
 - ソフトウェアルータ: JRouter
 - 接続用ソケット: JRServerSocket, JRSocket
 - 入出カストリーム: JROutputStream, JRInputStream
 - 管理クライアント: JRMonitor





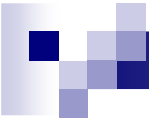
ソフトウェアルータ - JRouter

- 接続には2本のTCPストリームを使用
 - リレーデータ用と制御パケット用
 - 現状では受信バッファに空きが無い時は空くまで受信を待機
- Java New I/Oで複数のネットワーク入出力を管理
 - 単ースレッド動作でスレッドコンテキストの切替コスト削減
- GSI (Grid Security Infrastructure)による認証
 - 隣接ノード間認証と通信ピア間認証
 - 認証トークンをリレーすることで遠隔ノード間の認証を可能に



接続用ソケット - JRServerSocket, JRSocket

- ノードがJRouterに接続する際に使用
 - 認証コンテキストはJRouterとの認証用と通信ピアとの認証用の2つ用意 (ホスト証明書/ユーザ証明書)
- 通信モードを変更することでSSL暗号化通信可能
- ServerSocket/Socketと同様のインタフェース



入出力ストリーム - JROutputStream, JRInputStream

- 通信ピア間の入出力ストリーム
- 出力ストリーム: JROutputStream
 - JRouterのヘッダを付加
 - データのSSL暗号化
- 入力ストリーム: JRInputStream
 - JRouterでリレーするためのヘッダの解析・除去
 - SSLの復号化
- OutputStream/InputStreamを継承

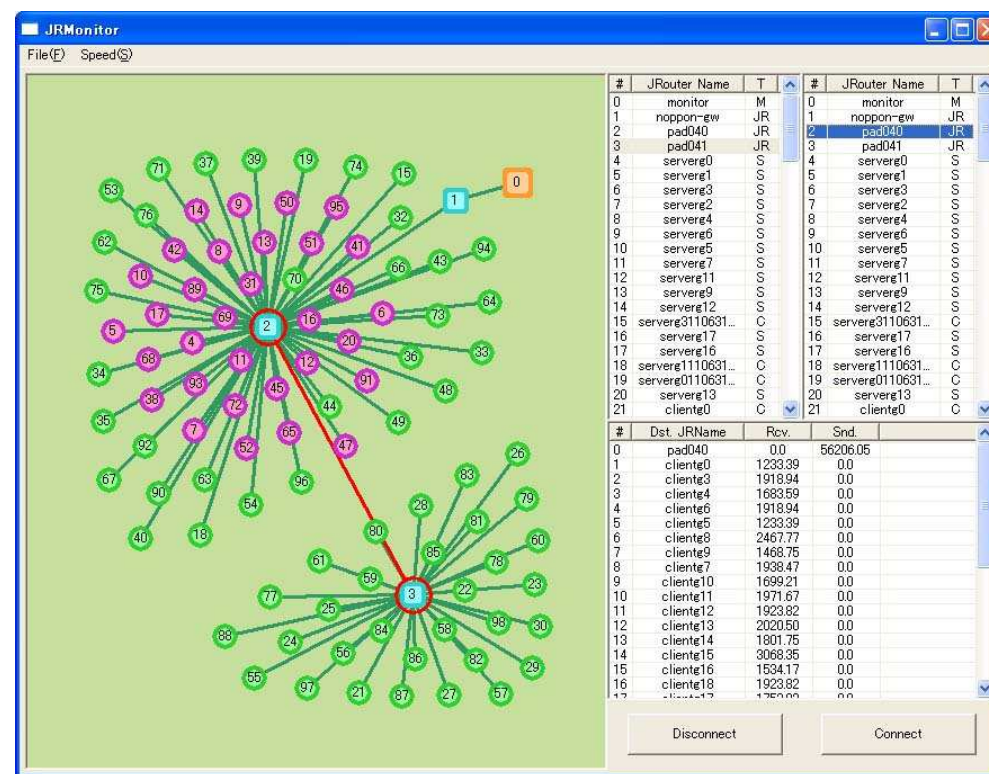
管理クライアント - JRMonitor

■ ネットワークトポロジの状態を表示

- トポロジの視覚化
- JRouterの通信状態

■ 管理機能

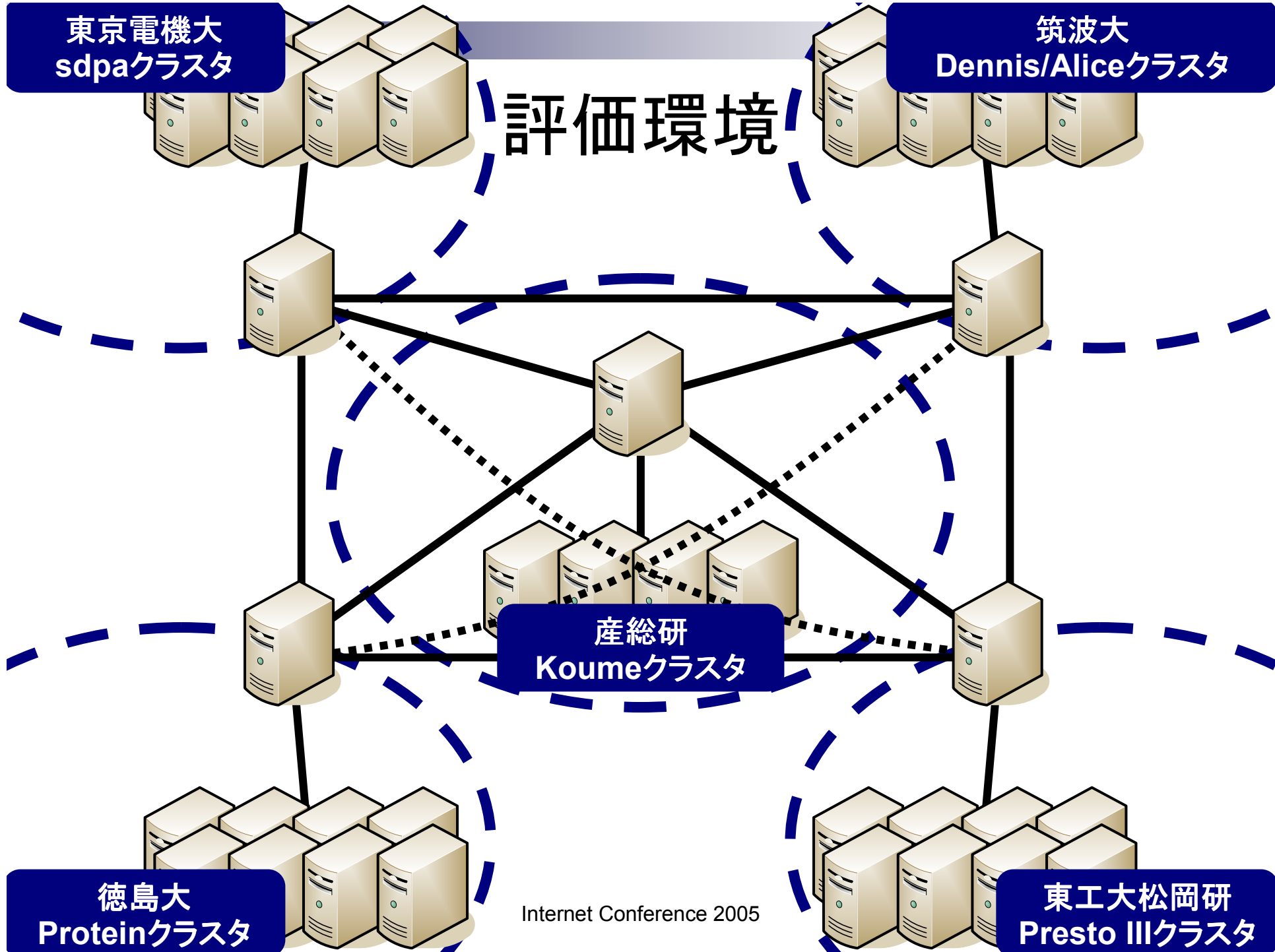
- リモート接続/切断





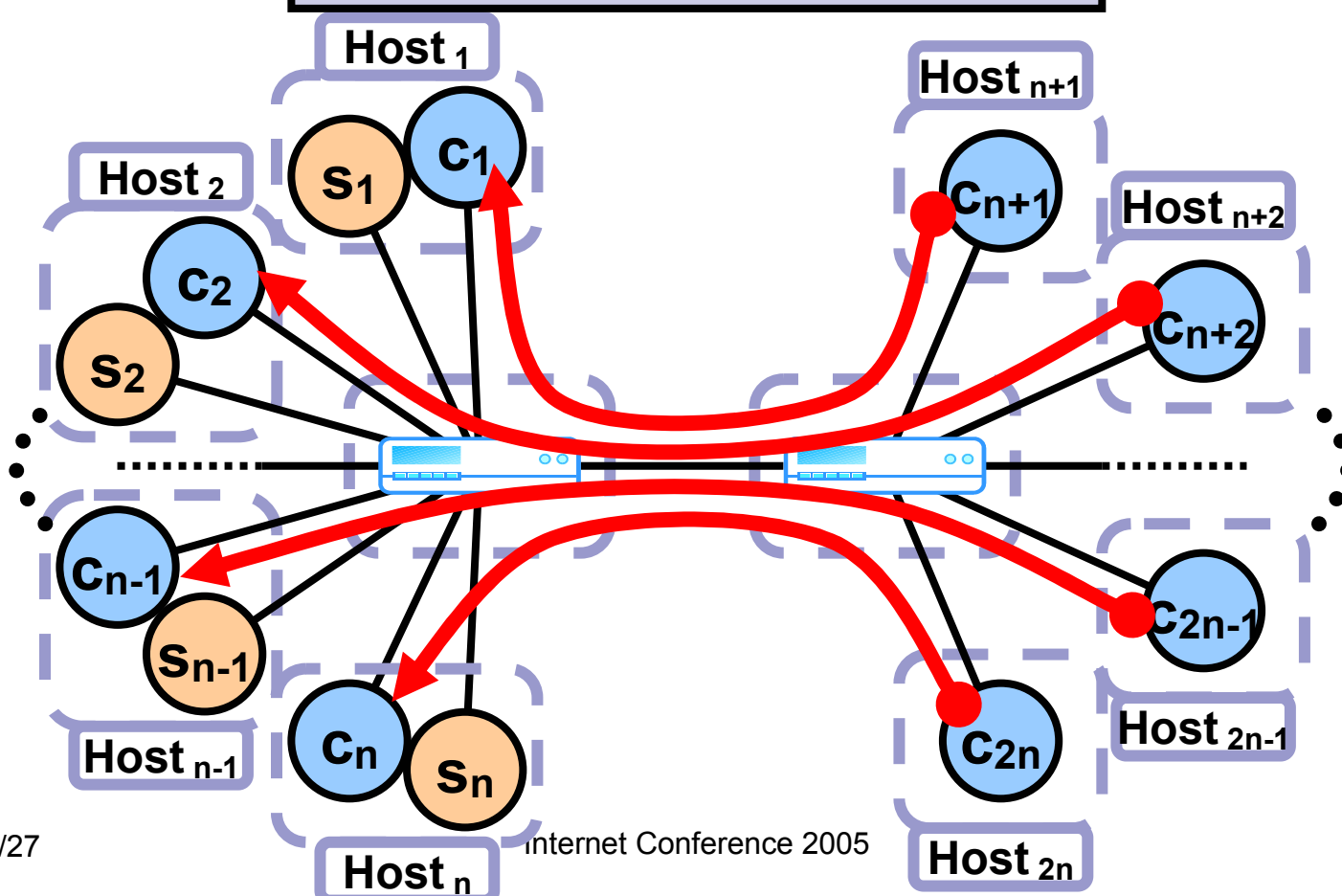
アジェンダ

- 背景・目的と要件定義
- 既存の隠蔽手法
- 新手法の提案と設計
- プロトタイプ実装
- **評価・考察**
- まとめ



基礎評価 - 2サイト間通信モデル

送信し続け、定常状態の時の
総スループットを計測



東京電機大
sdpaクラスタ

筑波大
Dennis/Aliceクラスタ

評価環境

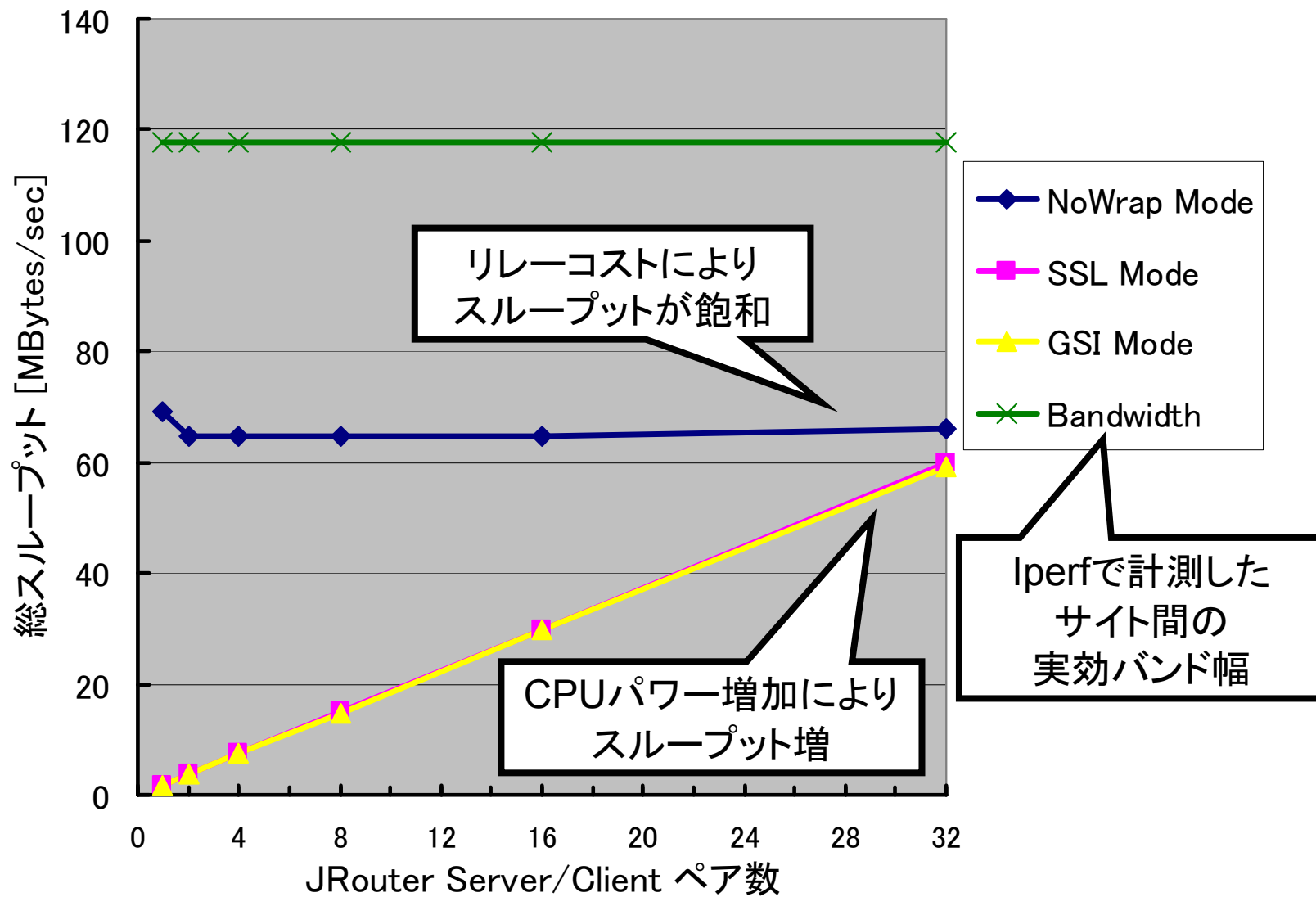
産総研
Koumeクラスタ

徳島大
Proteinクラスタ

東工大松岡研
Presto IIIクラスタ

CPU	Opteron242 x2
Mem	2GBytes
NIC	1000BASE-T
OS	Linux 2.4.27

2サイト間通信モデルによるスループット計測 PrestoIII内



東京電機大
sdpaクラスタ

筑波大
Dennis/Aliceクラスタ

評価環境

CPU	Xeon 2.4GHz x2	Athlon 1800+ x2
Mem	1GBytes	1.5GBytes
NIC	1000BASE-T	100BASE-TX
OS	Linux 2.4.20	Linux 2.4.19

CPU	Xeon 2.4GHz x2
Mem	1GBytes
NIC	1000BASE-T
OS	Linux 2.4.20

CPU	Opteron242 x2
Mem	2GBytes
NIC	1000BASE-T
OS	Linux 2.4.27

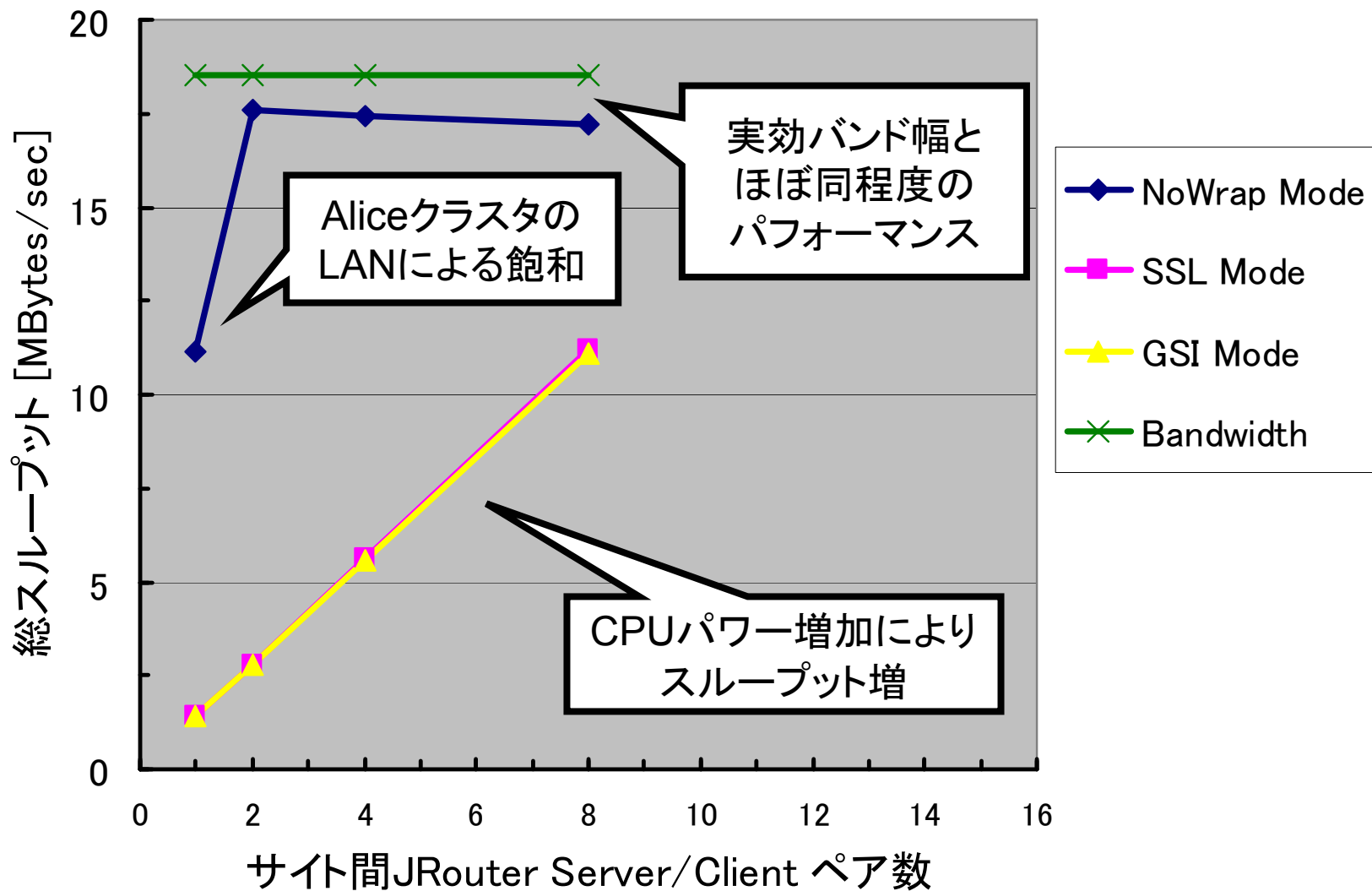
CPU	Opteron242 x2
Mem	2GBytes
NIC	1000BASE-T
OS	Linux 2.4.27

産総研
Koumeクラスタ

徳島大
Proteinクラスタ

東工大松岡研
Presto IIIクラスタ

2サイト間通信モデルによるスループット計測 (Aliceクラスター→Presto IIIクラスター)



東京電機大
sdpa クラスタ

CPU	Athlon MP 2400+ x2
Mem	1GBytes
NIC	1000BASE-T
OS	Linux 2.4.21

筑波大
Dennis/Alice クラスタ

評価環境

CPU	Athlon MP 1.33GHz
Mem	768MBytes
NIC	1000BASE-T
OS	Linux 2.4.22

CPU	Athlon MP 2000+ x2
Mem	512MBytes
NIC	100BASE-TX
OS	Linux 2.4.18

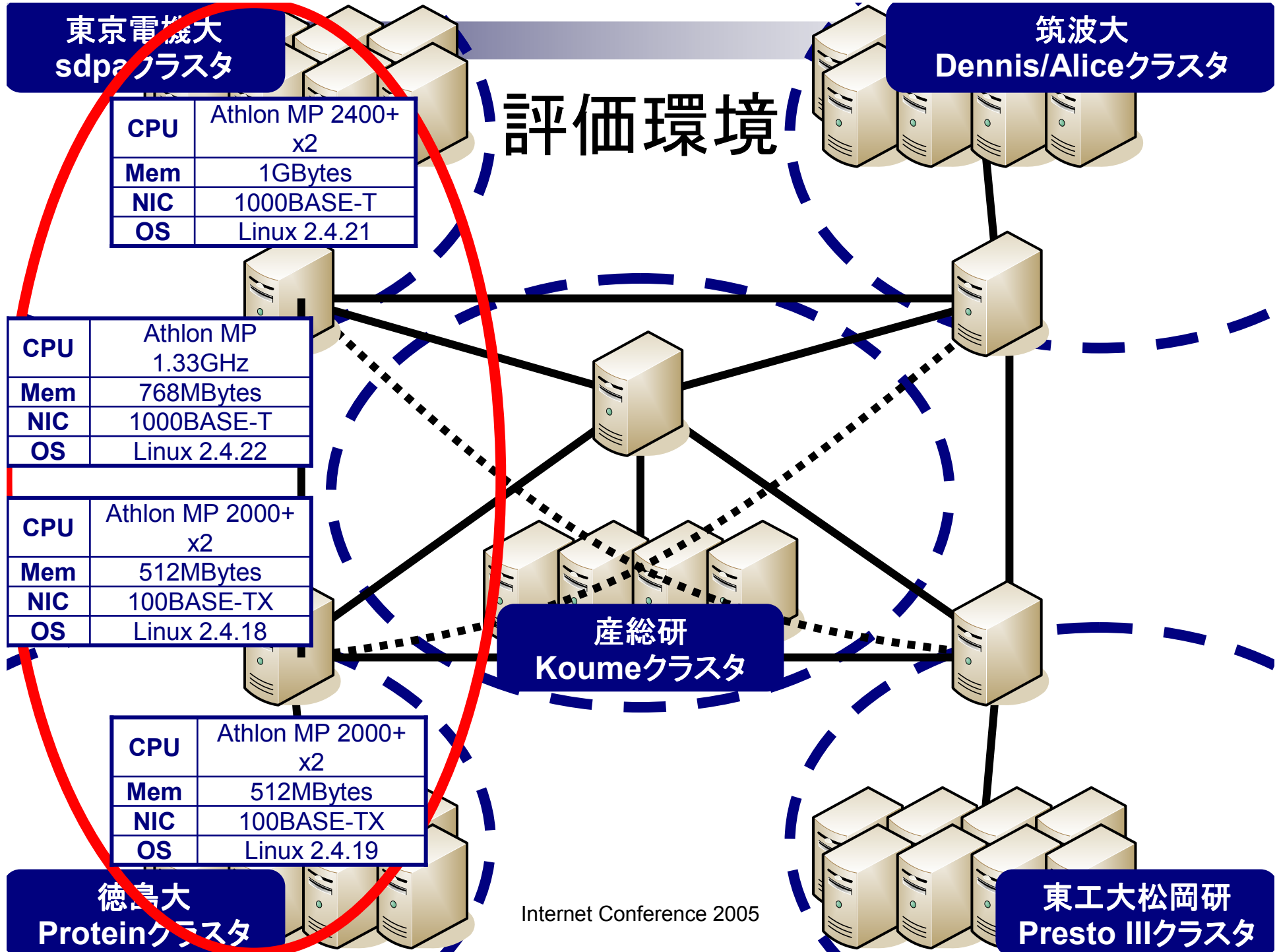
産総研
Koume クラスタ

CPU	Athlon MP 2000+ x2
Mem	512MBytes
NIC	100BASE-TX
OS	Linux 2.4.19

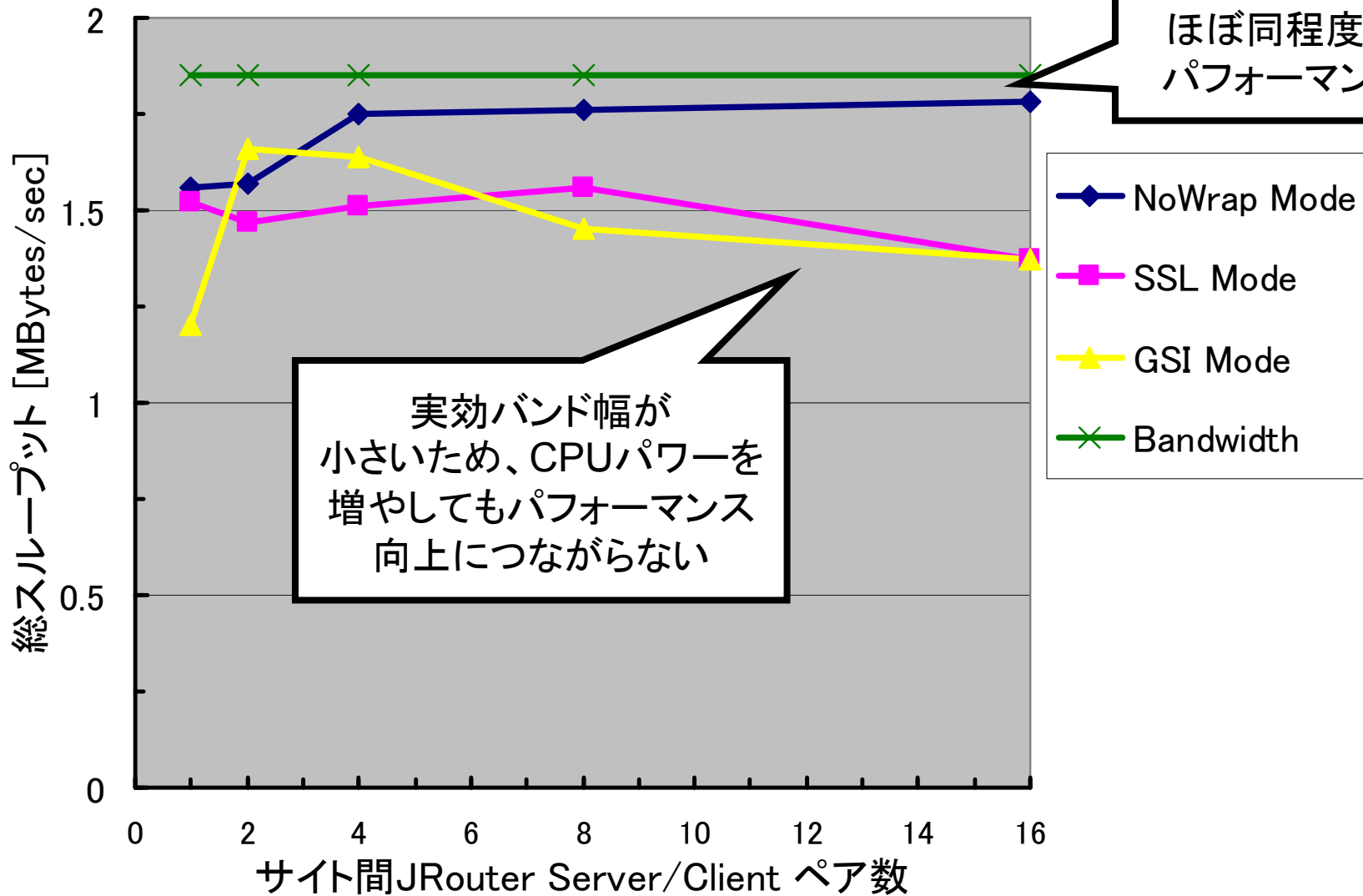
徳島大
Protein クラスタ

Internet Conference 2005

東工大松岡研
Presto III クラスタ



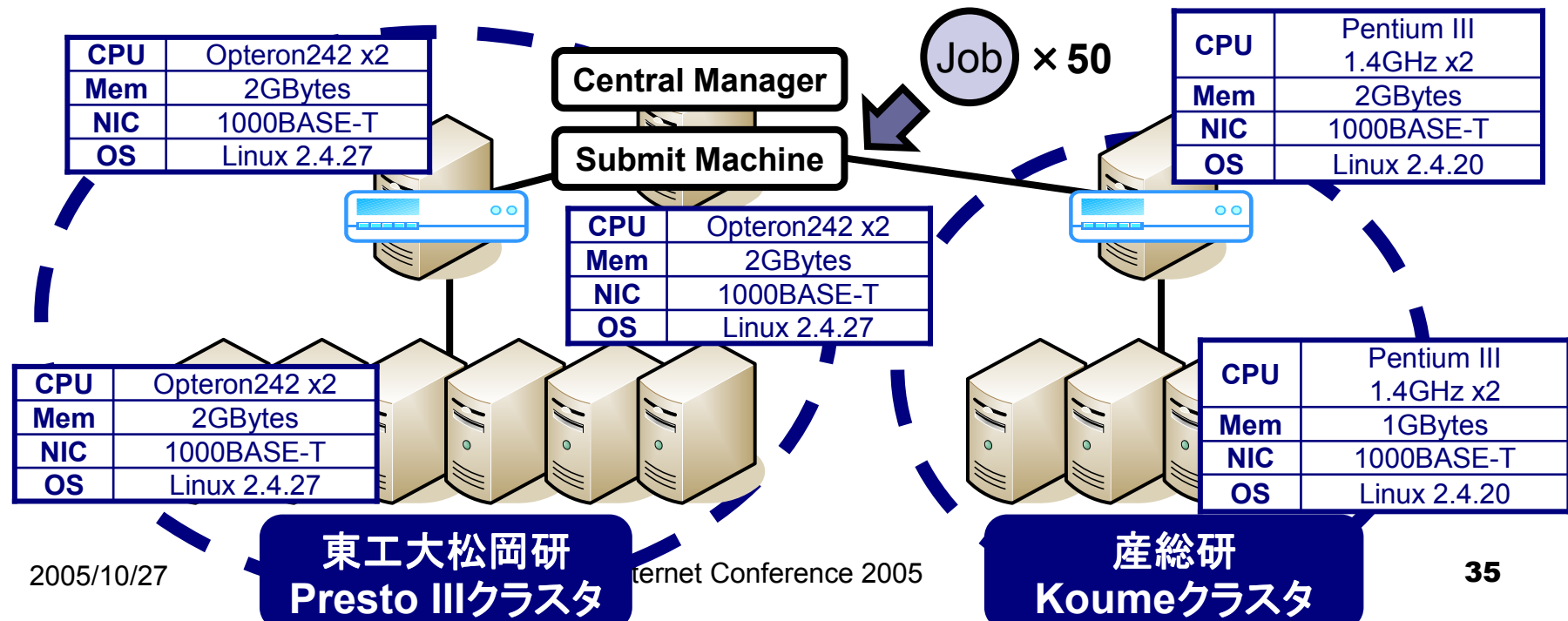
2サイト間通信モデルによるスループット計測 sdpaクラスター→Proteinクラスター



実アプリケーションによる評価

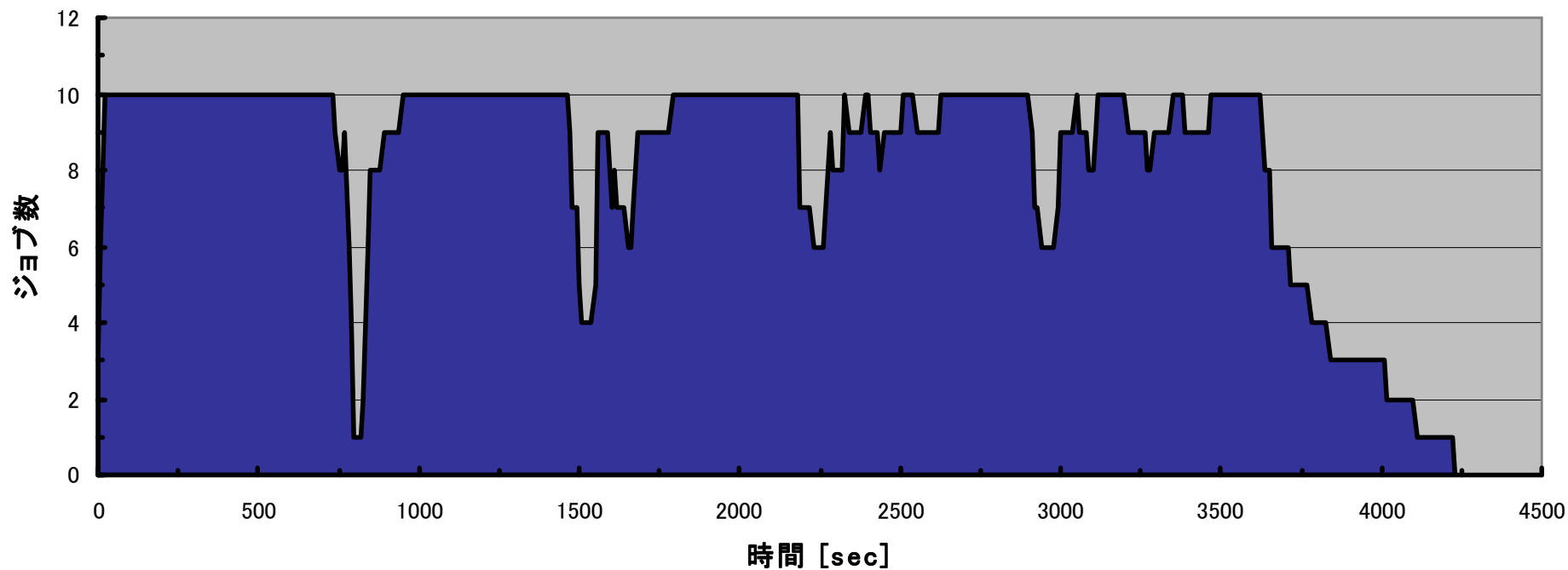
■ 評価環境

- 耐故障性に優れたジョブスケジューリングシステムJay
- ジョブにはホモロジー検索プログラムblastを使用





Jayによるジョブ起動数の変化



実際に非対称ネットワークが問題となっていた
ジョブスケジューリングシステムでの稼働を確認




考察

- 接続性
 - 異なるプライベート空間のホスト間で通信可能
- セキュリティ
 - 通信ピア間での認証・認可・暗号化
- サイトポリシ非依存
 - Super User権限の無いサイトでの動作を確認
 - 異なる管理ポリシを持つ5サイトで動作確認
- 通信性能
 - 実効バンド幅の小さいWAN環境ではリレーコストによる性能低下は見られない
 - 実効バンド幅の大きいローカルサイトで通信性能低下



更なる高通信性能に向けて

- JRouterの受信バッファを増加
- CPUパワーの余剰で通信データの圧縮
- 通信プロトコルの見直し
- リアルタイムスループット計測に基づくマルチパス転送



アジェンダ

- 背景・目的と要件定義
- 既存の隠蔽手法
- 新手法の提案と設計
- プロトタイプ実装
- 評価・考察
- **まとめ**



まとめ

- 非対称ネットワークを隠蔽する高速通信インフラストラクチャを提案
- プロトタイプであるJRouterの実装と評価
- JRouterが接続性・セキュリティ・サイトポリシー非依存性を満たすこと確認
- 更なる高通信性能のための対策を考察