ヤングリサーチャー発表枠

大規模機械学習向けクラスタにおける ネットワーク構造とパラメータ交換手法

黎 明曦（筑波大）
谷村勇輔（産総研・筑波大）
◯中田 秀基（産総研・筑波大）

研究背景

- 機械学習：大量のデータの処理
  - 並列化による高速化が必要
    - モデル並列(model parallel)
    - データ並列(data parallel)

- データ並列機械学習システムの問題
  - 大規模並列機械学習向け計算システムのネットワークへの要請があきらかでない
  - ネットワークコストと計算への影響のトレードオフ

研究の概要

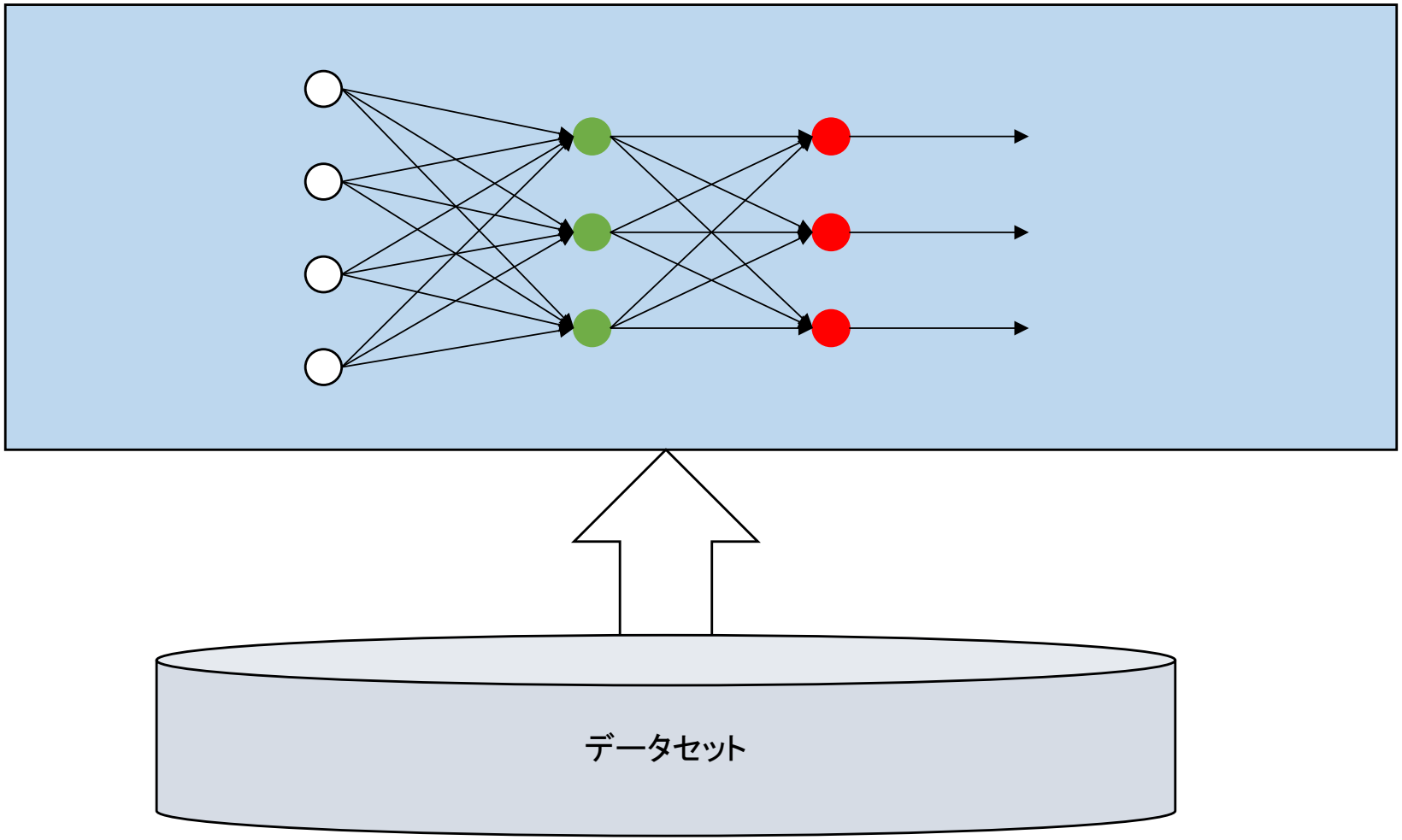
- バイセクションバンド幅とパラメータ交換手法の関係を調査
  - 一般的な2層ネットワークを前提に
  - 分散環境シミュレータSimGridを利用
  - いくつかのパラメータ交換手法を評価

発表の概要

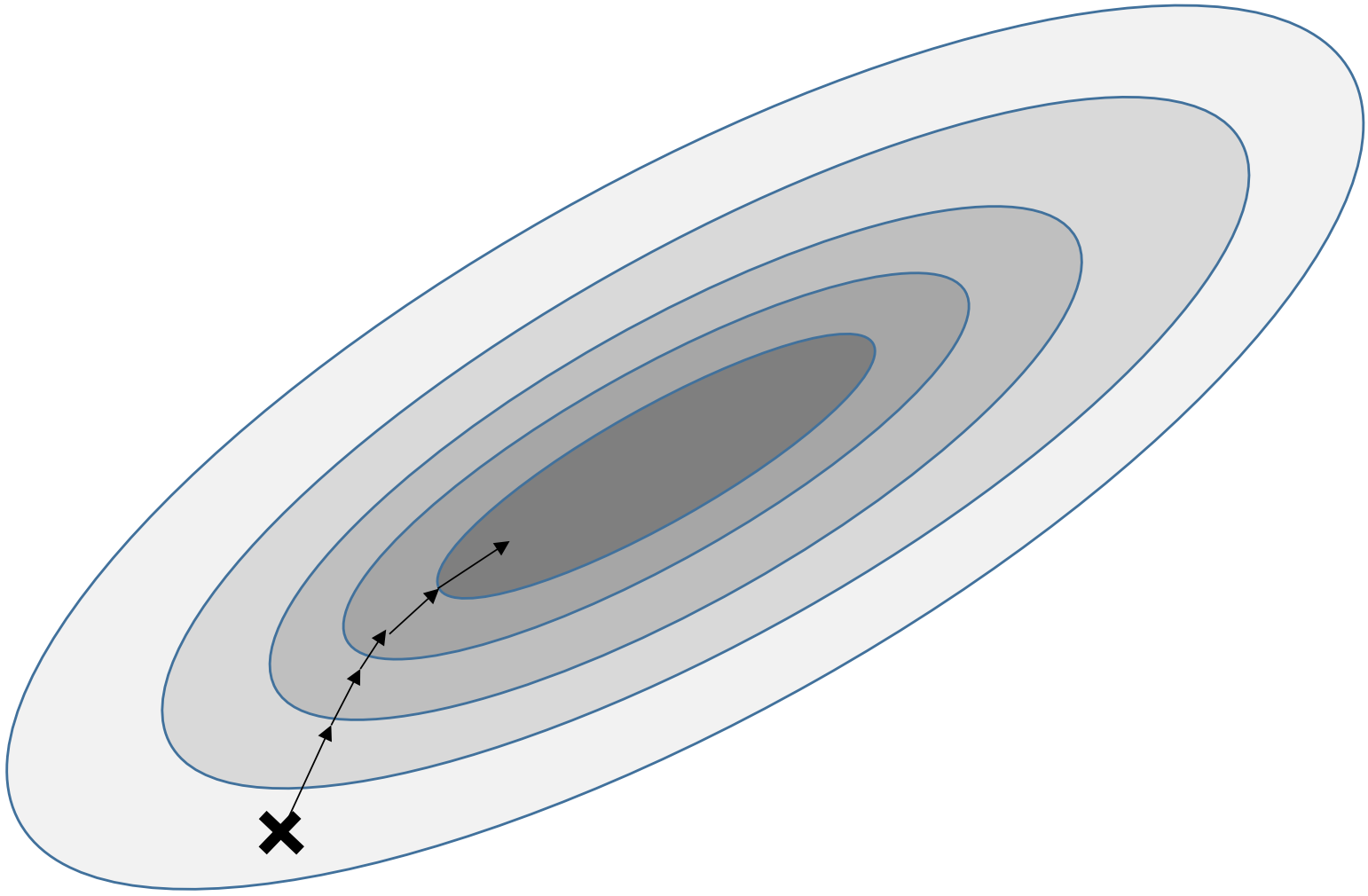
- 研究背景と概要
- 背景
  - データ並列機械学習
  - ネットワーク
  - SimGrid
- パラメータ交換手法
- 評価
  - 設定
  - 結果
- 関連研究
- 結論

機械学習の基本

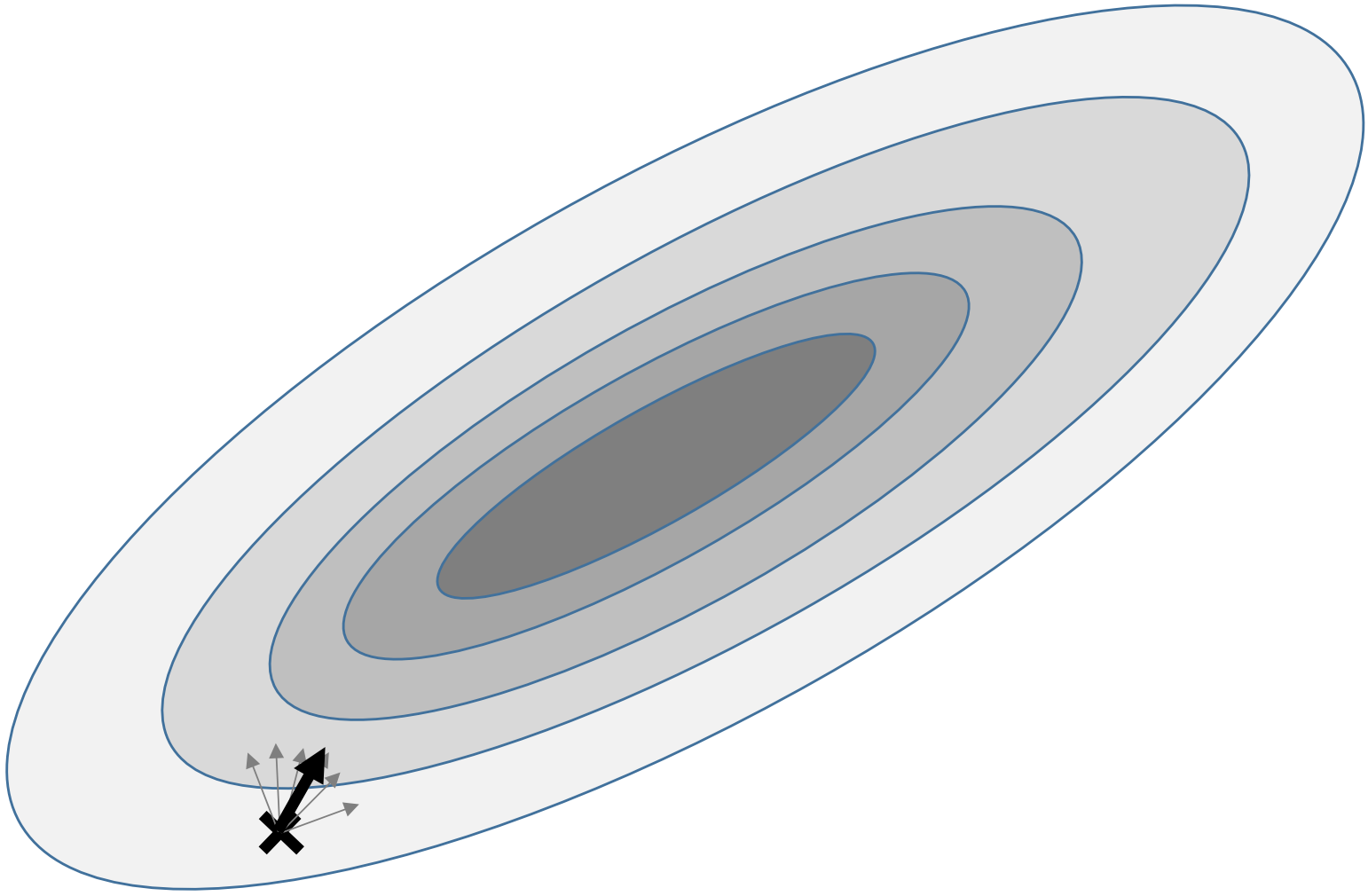
- 訓練用データセットを用いてパラメータを更新



機械学習の基本

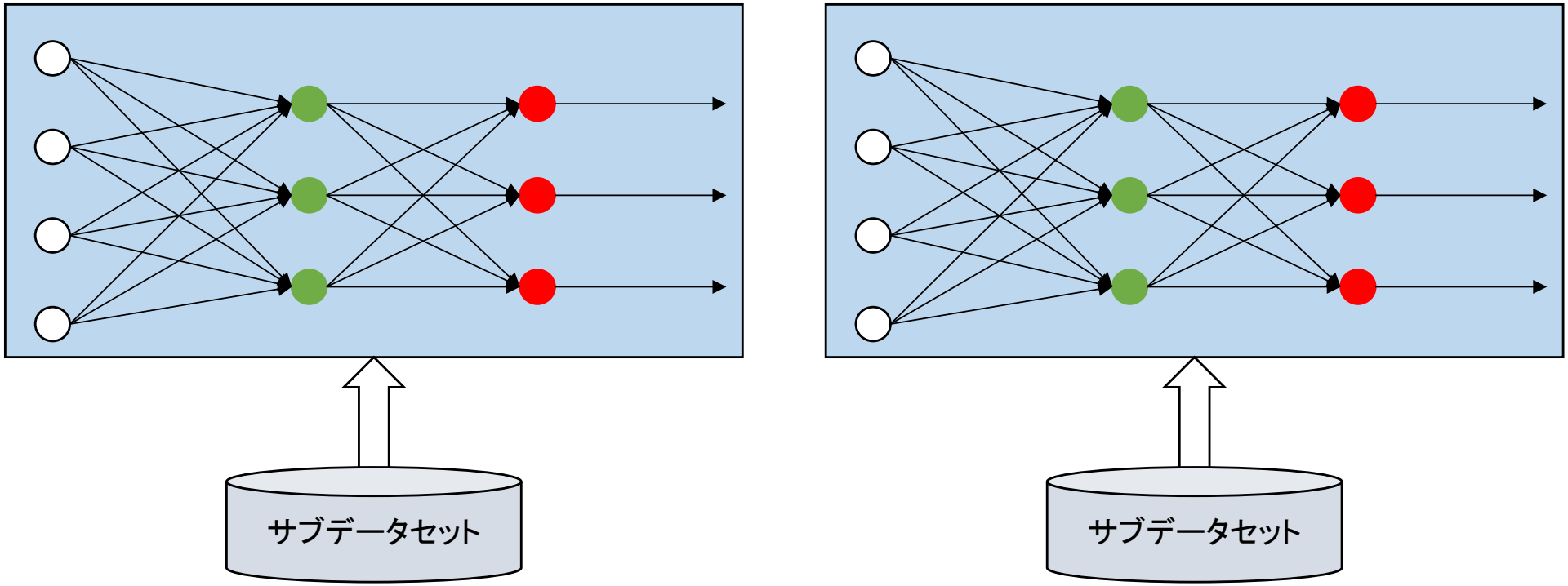


機械学習の基本



データ並列学習システム

- データセットを各機械学習機に割り当て
- グラディエントを定期的に交換



発表の概要

- 研究背景と概要

- 背景
  - データ並列機械学習
  - ネットワーク
  - SimGrid

- パラメータ交換手法

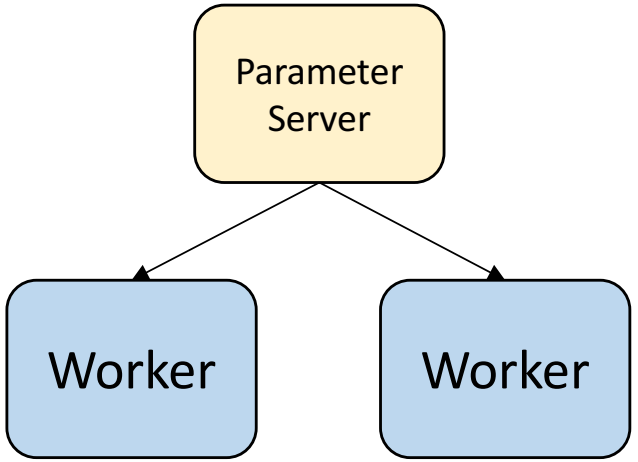
- 評価
  - 設定
  - 結果

- 関連研究

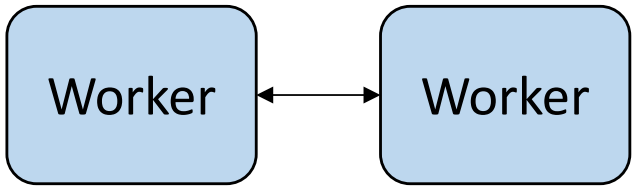
- 結論

パラメータ交換の方法

- パラメータサーバ法
  - 実装が楽
  - 非同期型が実装しやすい
  - FTに適する



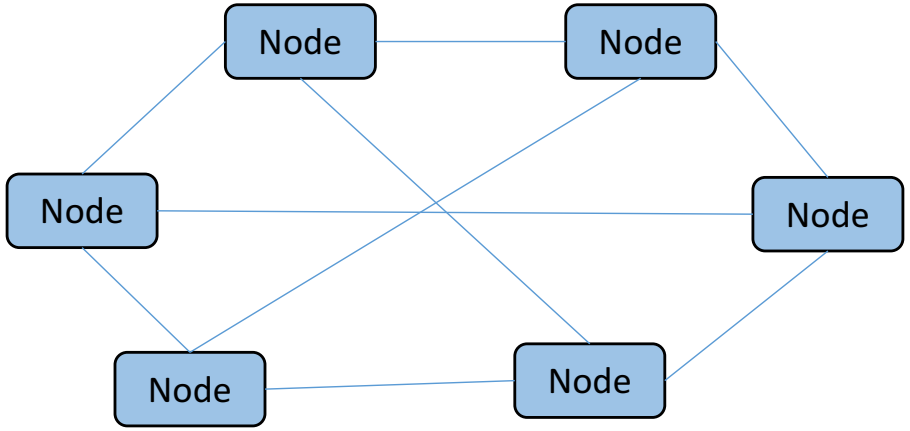
- 直接交換法
  - 実装はちょっと面倒
  - 非同期型にするのは大変



バイセクションバンド幅

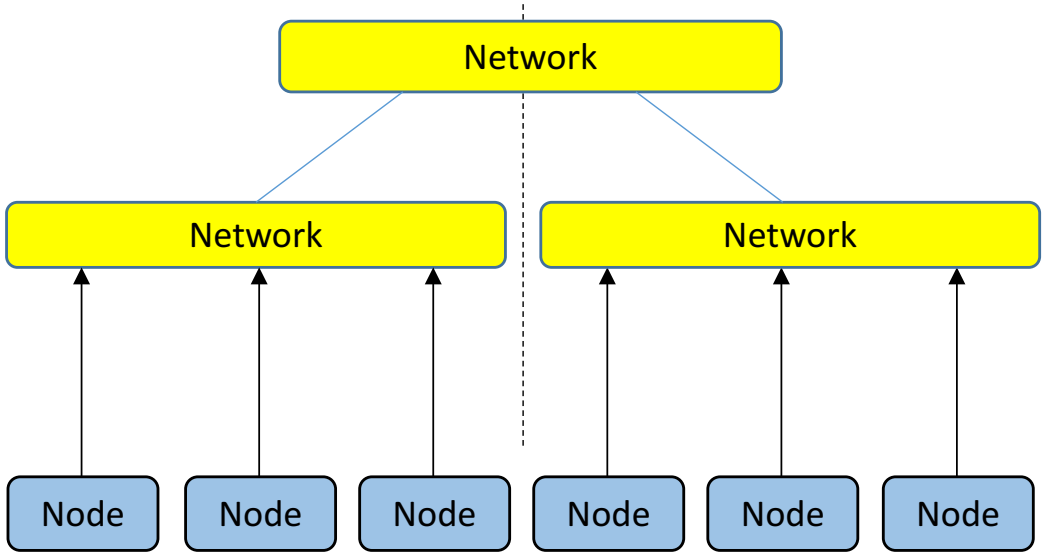
- バイセクションバンド幅

ネットワークを左右二等分した場合、双方をつなぐリンクのバンド幅の合計



- フルバイセクション

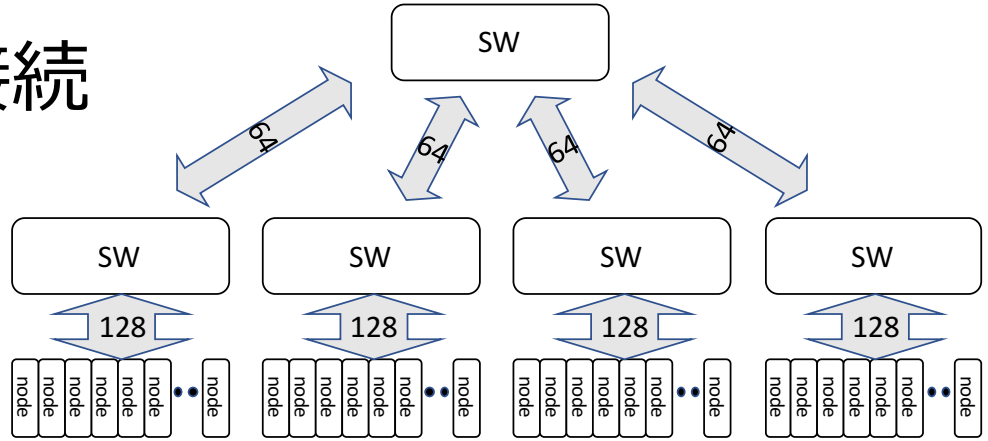
バイセクションバンド幅がネットワークの一方のバンド幅と同じ



ネットワークモデル

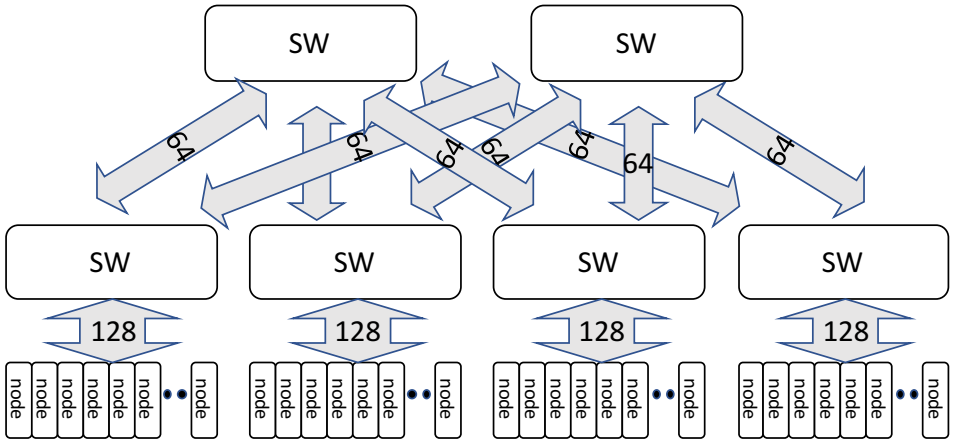
- 各クラスタ：128ノード
- 各スイッチ：最大256接続

- 比較的大規模なスイッチで構成した複数のサブクラスタを、上位のスイッチで接続してスケールアップ



ハーフバイセクション

- 上位のスイッチを複数設けてファットツリーを構成



フルバイセクション

SIMGRID

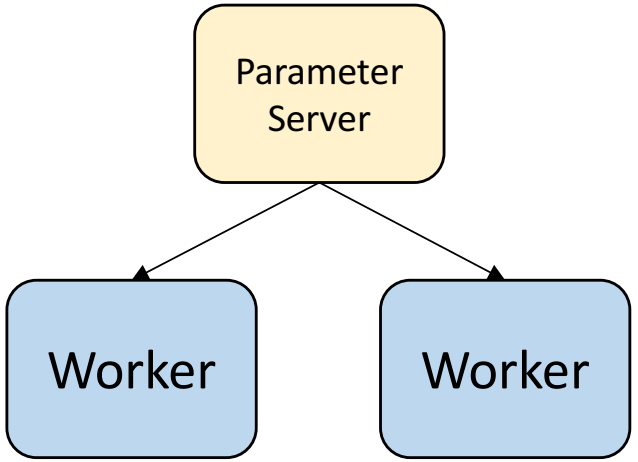
- シミュレーションフレームワーク
  - 1998年リリース
  - 開発 : UCSD
  - メンテナンス : Inria

- アプリケーションシミュレータ
  - イベントの実際発生 ❌
  - イベントのコストだけ抽出 ✔

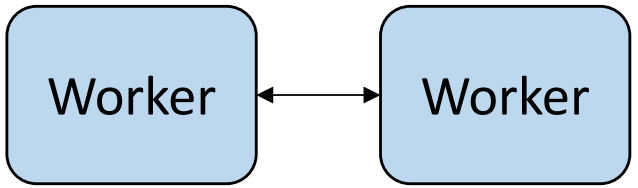
[1] Henri Casanova, Arnaud Giersch, Arnaud Legrand, Martin Quinson, Frédéric Suter.
Versatile, Scalable, and Accurate Simulation of Distributed Applications and Platforms.
Journal of Parallel and Distributed Computing, Elsevier, 2014, 74 (10), pp.2899-2917.

パラメータ交換の方法

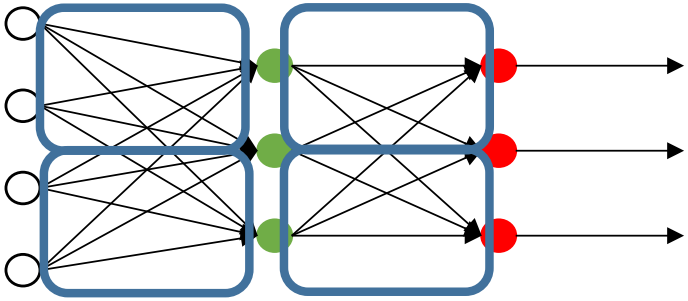
- パラメータサーバ法
  - 実装が楽
  - 非同期型が実装しやすい
  - FTに適する

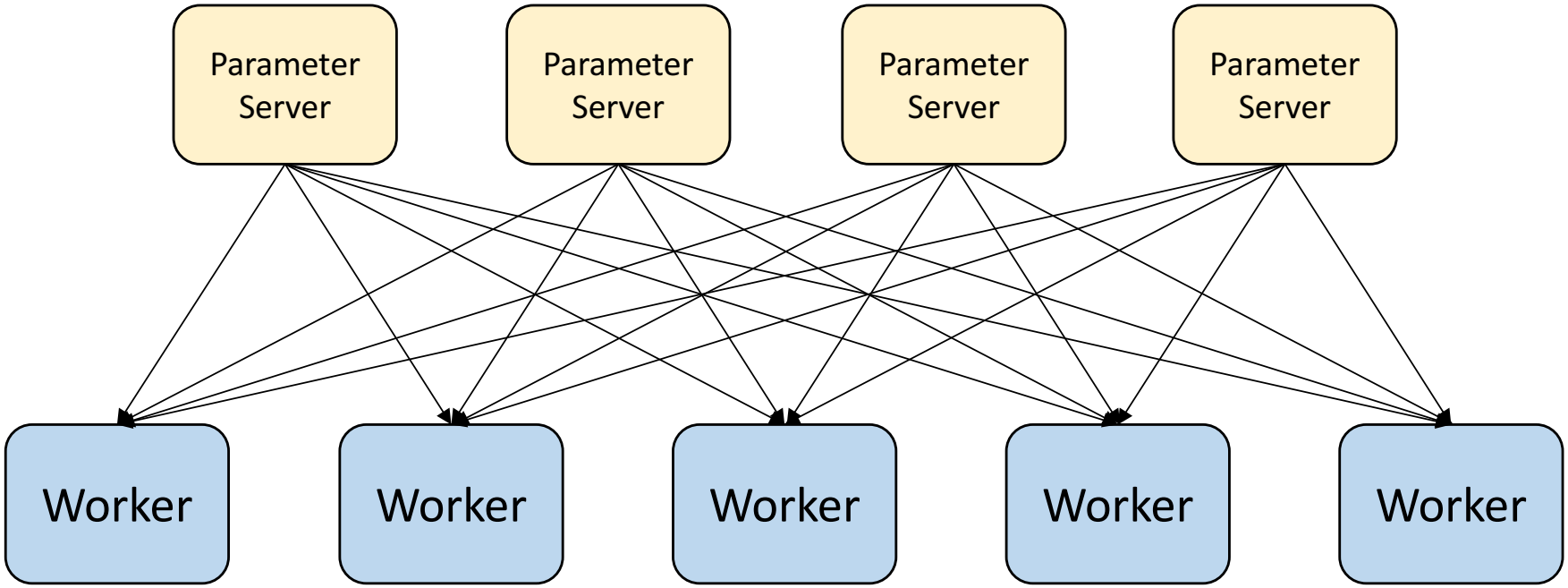


- 直接交換法
  - 実装はちょっと面倒
  - 非同期型にするのは大変



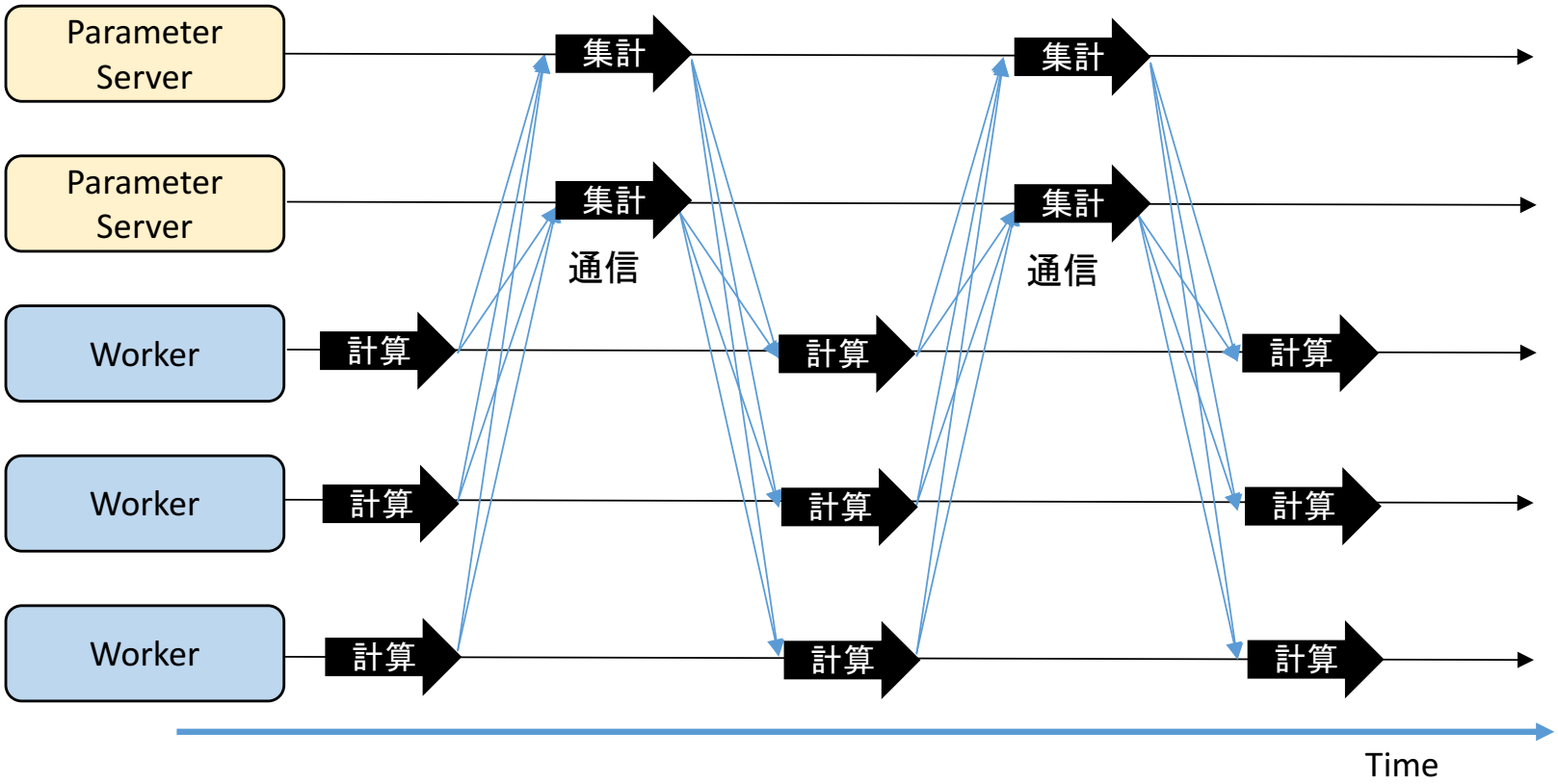
パラメータサーバとは




Parameter Server
Parameter Server
Parameter Server
Parameter Server
Worker
Worker
Worker
Worker
Worker

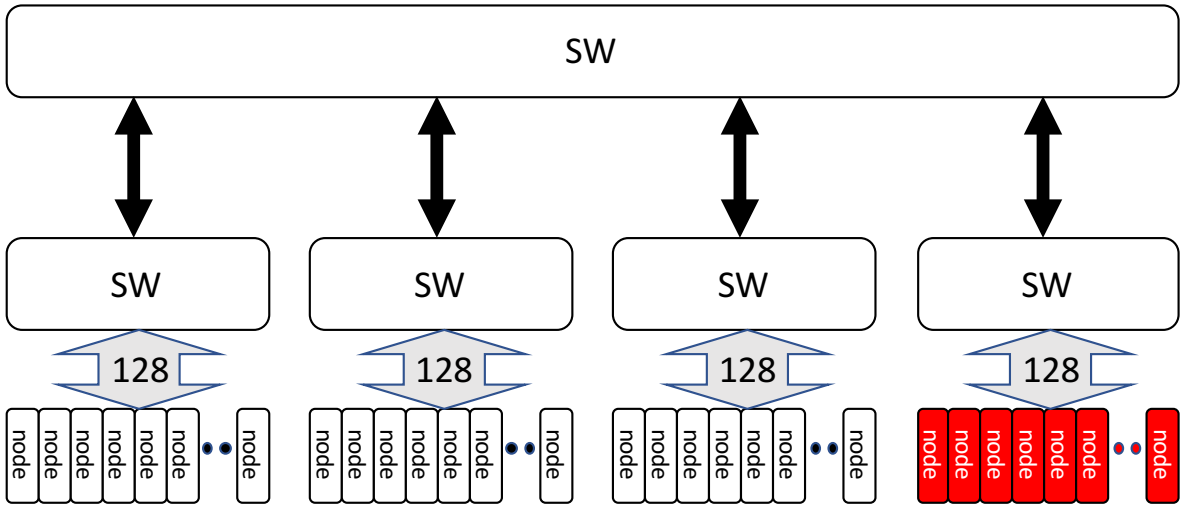
パラメータサーバの通信パターン

- サーバノードは複数：ボトルネック防止
  各サーバノードはパラメータの一部を扱う。

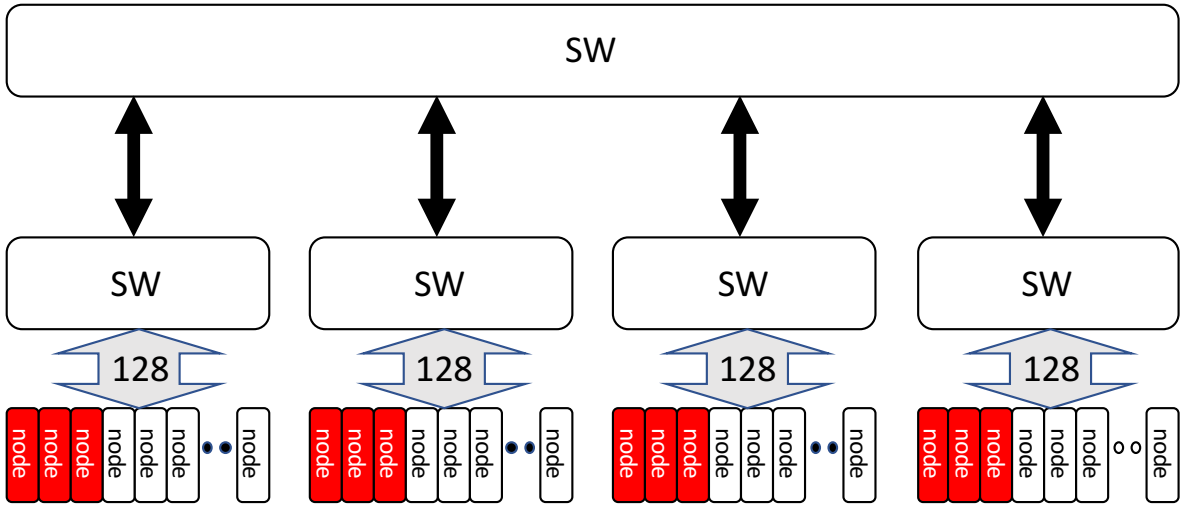


パラメータサーバによる同期

- 集中サーバ
  - 一つのサブクラスタをすべてパラメータサーバに



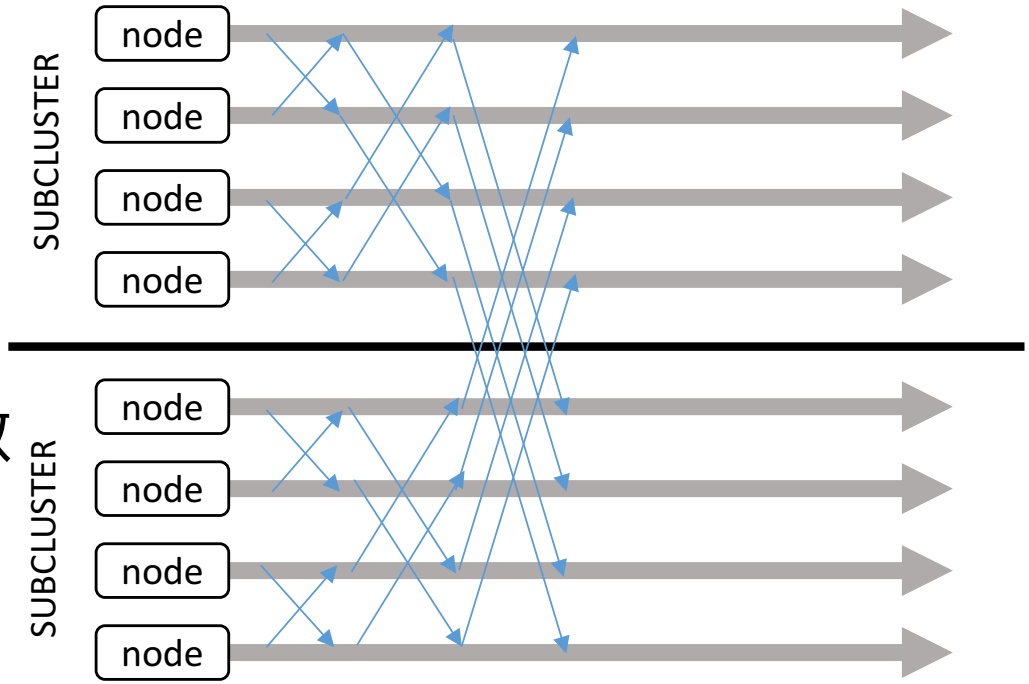
- 分散サーバ
  - パラメータサーバをサブクラスタに均等に分散



直接パラメータ交換

- 直接パラメータ交換
  - バタフライ
  - 二層バタフライ

通信段数：Log_2NM
N ＝ クラスタ内ノード数
M ＝ サブクラスタ数



直接パラメータ交換

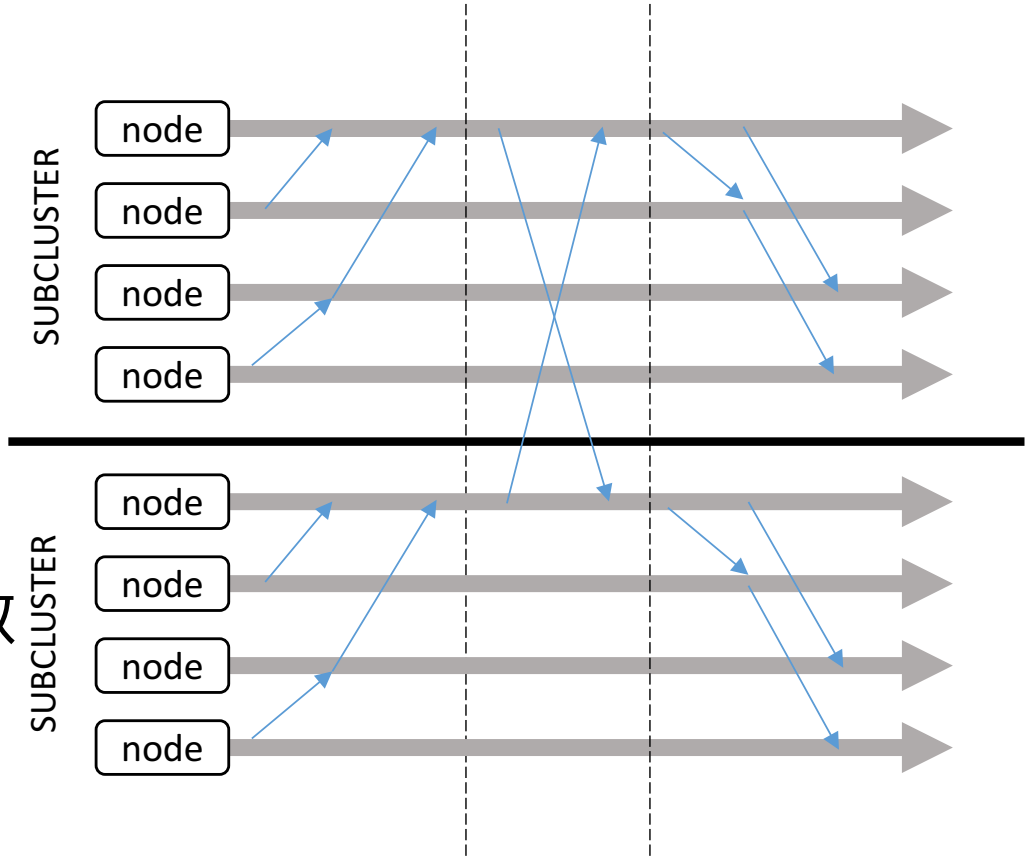
- 直接パラメータ交換
  - 普通のバタフライ
  - 二層バタフライ

通信段数:

$2\mathrm{Log}_2N+\mathrm{Log}_2M$

N = クラスタ内ノード数

M = サブクラスタ数



発表の概要

- 研究背景と概要

- 背景
  - データ並列機械学習
  - ネットワーク
  - SimGrid

- パラメータ交換手法

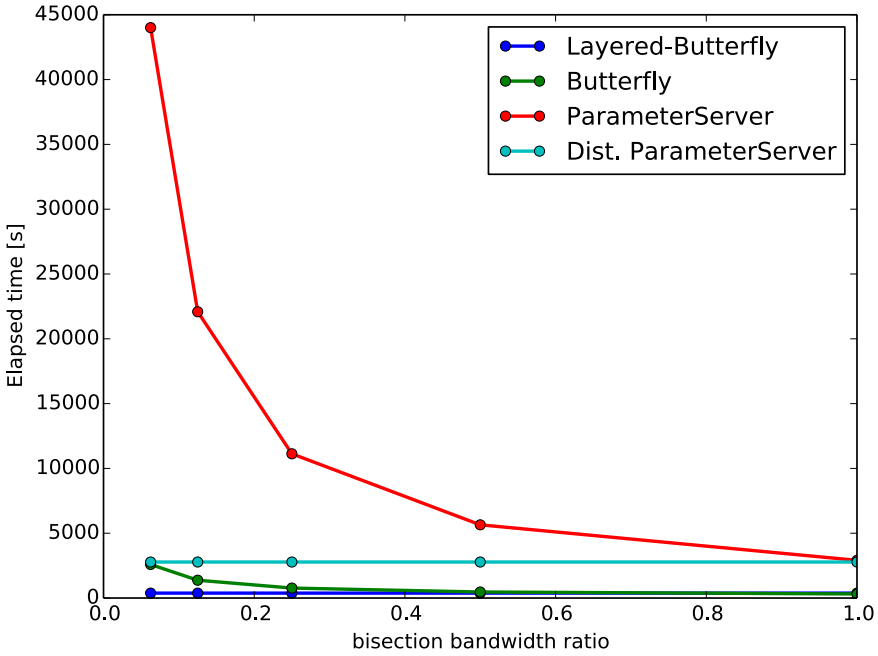
- 評価
  - 設定
  - 結果

- 関連研究

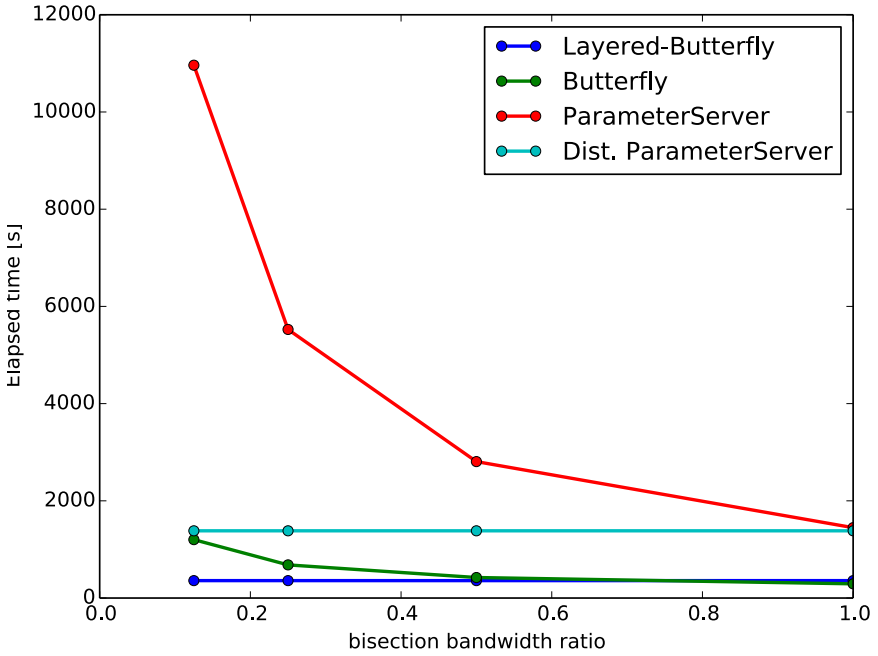
- 結論

評価

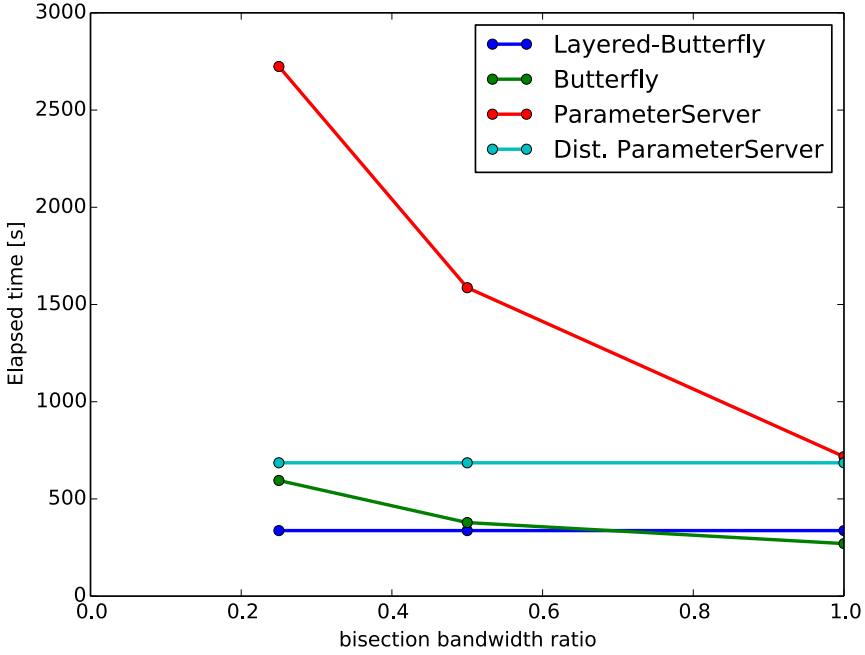
- サブクラスタ数：16、32、64、128
- サブクラスタのノード数：128
- パラメータのサイズ：1GB
- パラメータ交換間隔（1回学習時間）：1s
- パラメータ交換回数：10回
- バイセクションバンド幅：フル、ハーフ、1/4、1/8…



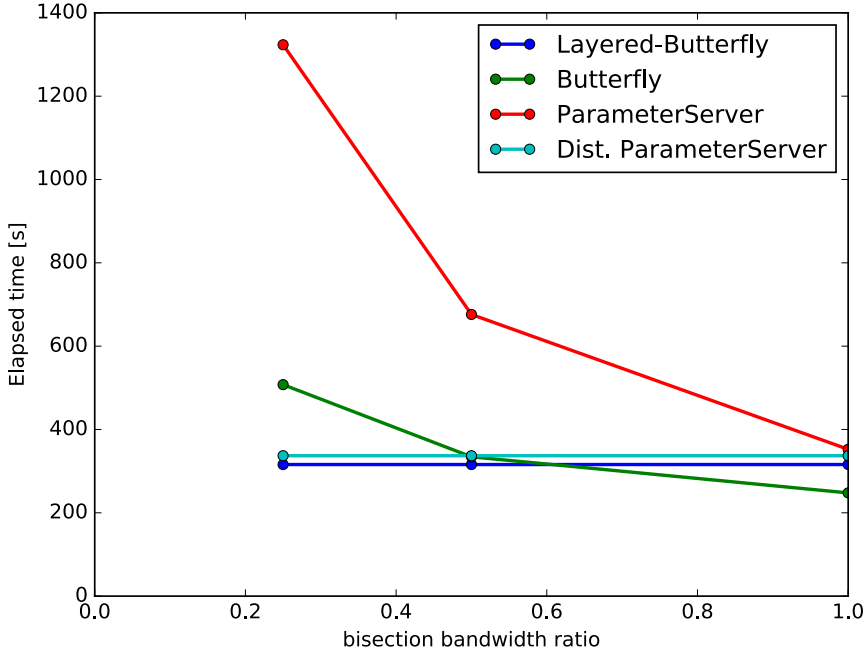
128サブクラスタ



64サブクラスタ



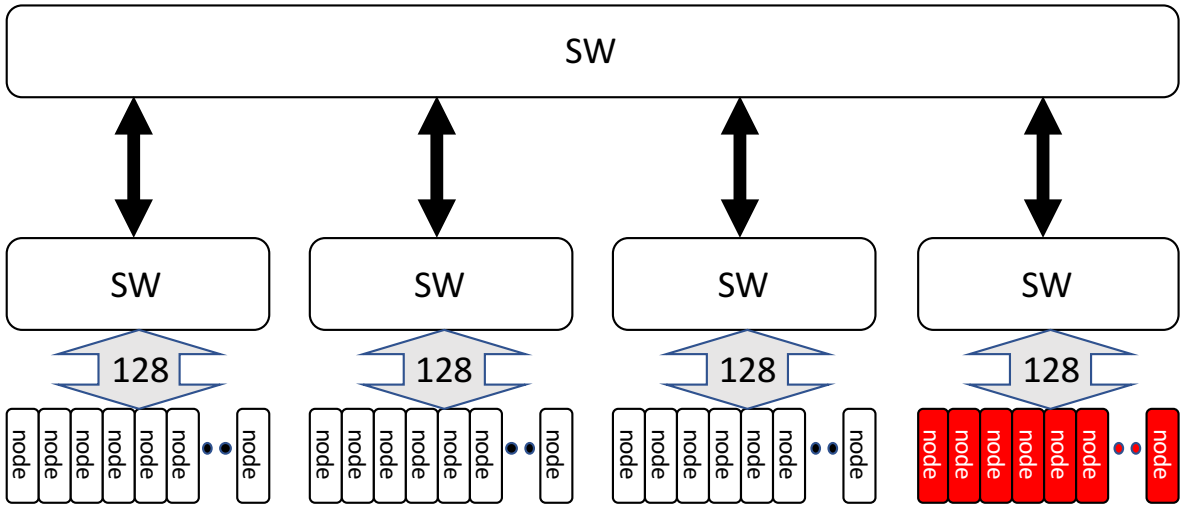
32サブクラスタ



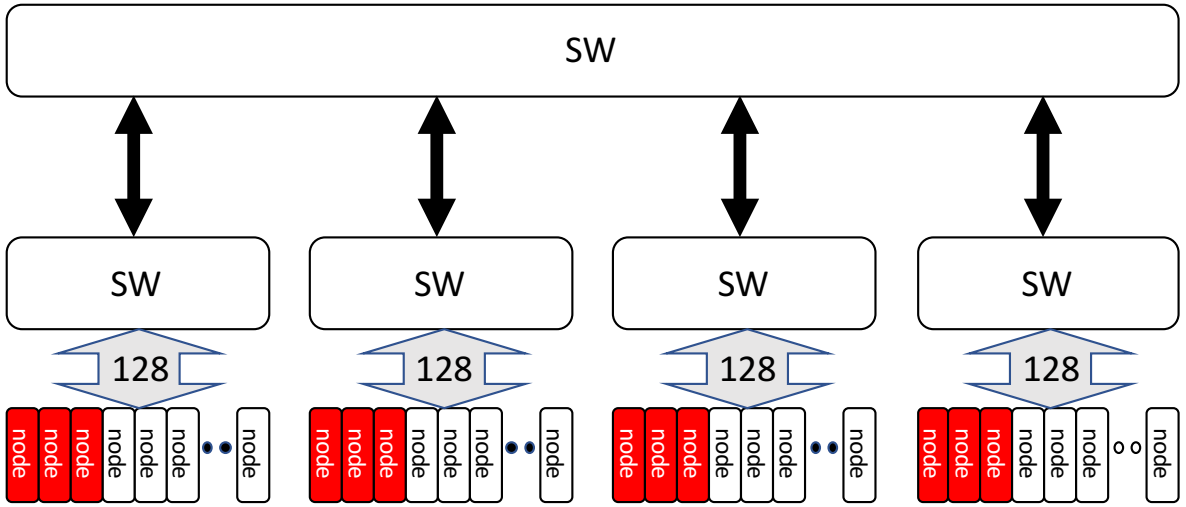
16サブクラスタ

パラメータサーバによる同期

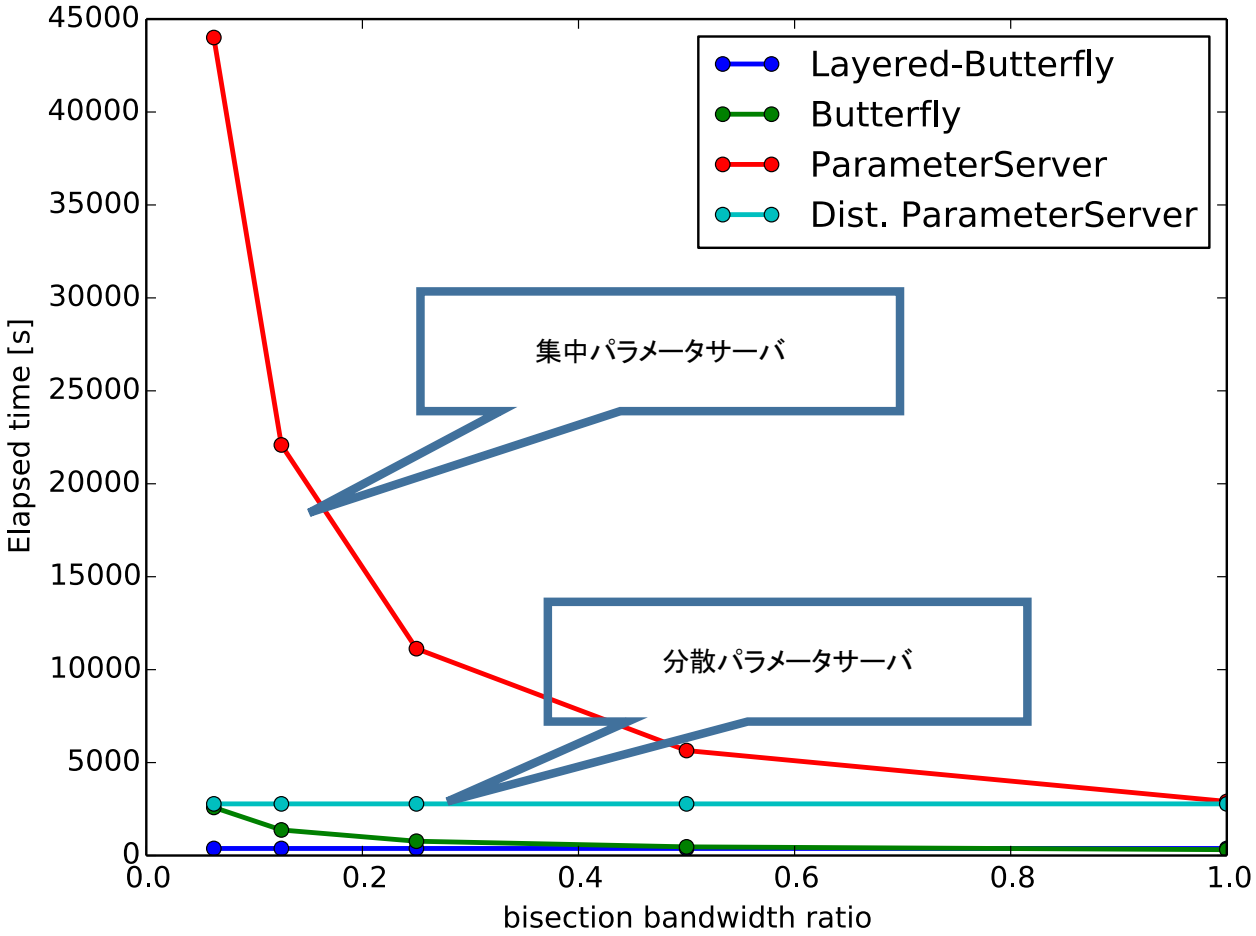
- 集中サーバ
  - 一つのサブクラスタをすべてパラメータサーバに



- 分散サーバ
  - パラメータサーバをサブクラスタに均等に分散



パラメータサーバの評価結果

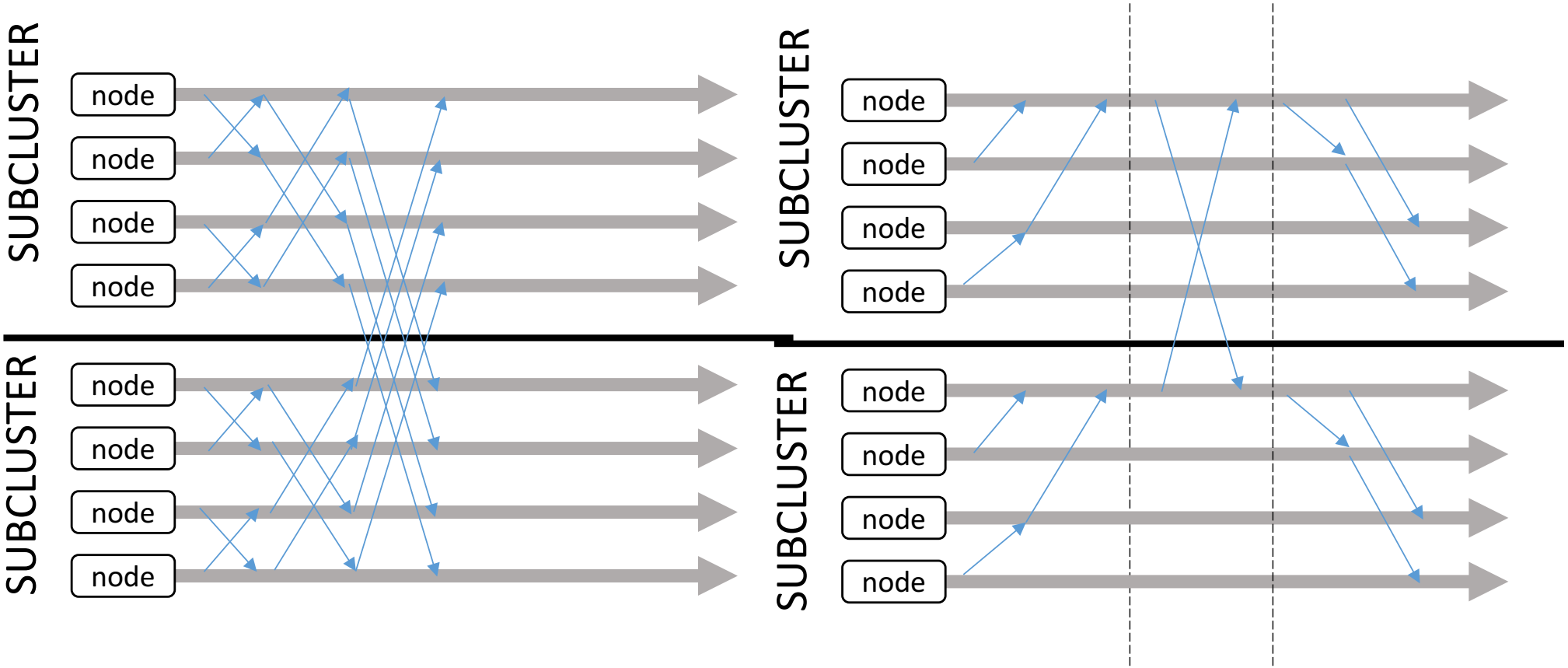


交換モデルサイズ1Gbyte
ネットワーク速度1GByte/s
パラメータ交換間隔1秒、10回の交換で測定

- パラメータサーバはバタフライと比較して一般に低速
- 十分なバンド幅があれば、集中型と分散型の性能のは同等
- 集中パラメータサーバは特にバイセクションバンド幅の低下に敏感

直接パラメータ交換

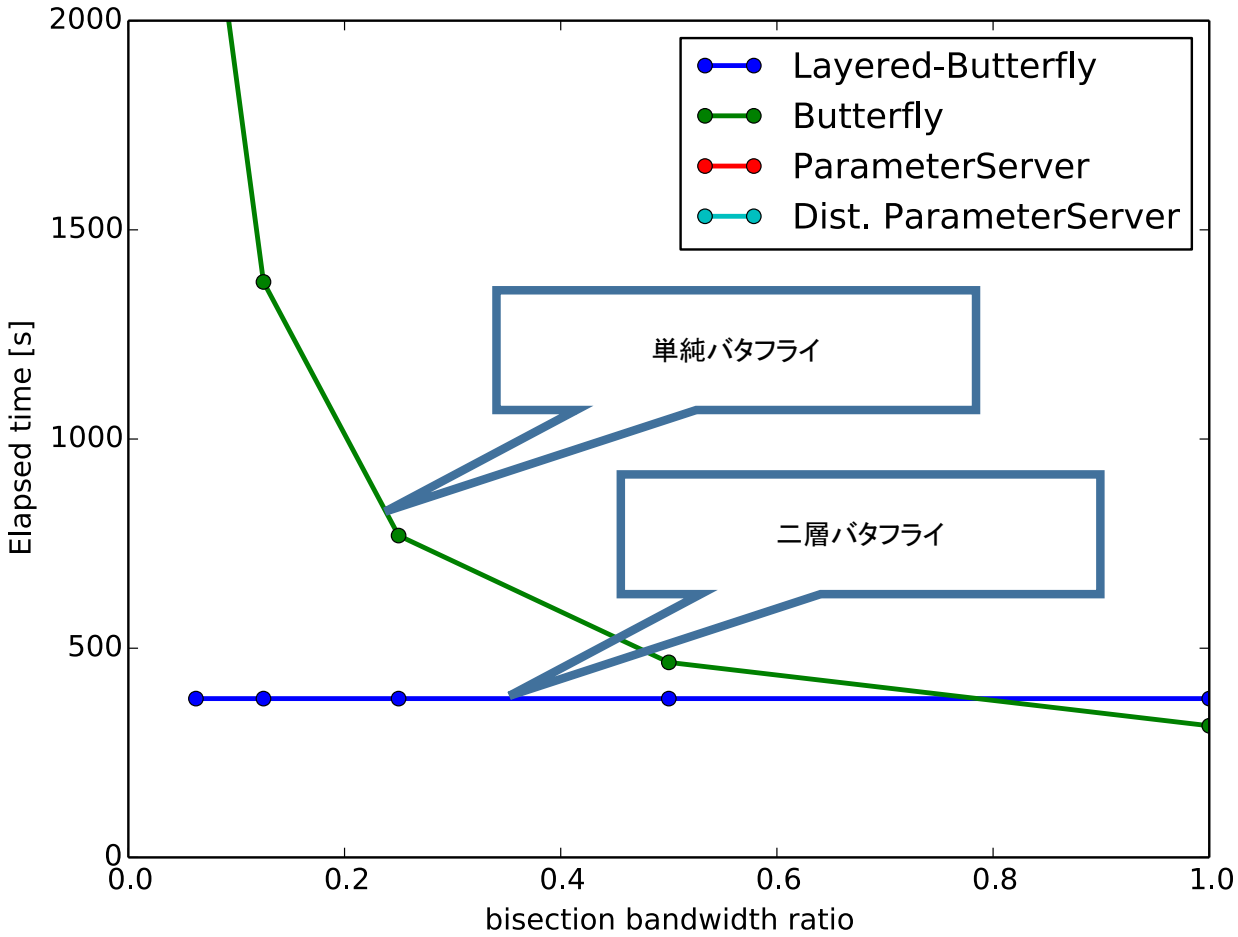
- 単純バタフライ
- 二層バタフライ



- 通信段数：Log2NM
  N = クラスタ内ノード数
  M = サブクラスタ数

- 通信段数:2Log2N+Log2M
  N = クラスタ内ノード数
  M = サブクラスタ数

直接パラメータ交換の評価結果



交換モデルサイズ1Gbyte
ネットワーク速度1GByte/s
パラメータ交換間隔1秒、10回の交換で測定

- 十分なバンド幅があれば、単純バタフライのほうが高速
- 単純バタフライはバイセクションバンド幅の低下に敏感
- 二層バタフライはバイセクションバンド幅の低下に影響を受けない

発表の概要

- 研究背景と概要
- 背景
  - データ並列機械学習
  - ネットワーク
  - SimGrid
- パラメータ交換手法
- 評価
  - 設定
  - 結果
- 関連研究
- 結論

関連研究

- パラメータ交換手法
  - 非同期パラメータ交換
    - 完全に同期しなくても計算は成立する
    - ただし、収束が遅くなる/精度が下がる
  - ゴシッププロトコル
    - ランダムな相手に自分のグラディエントを送るだけ

➔同期、非同期の動的な切り替え？

関連研究

- パラメータ交換手法
  - Gradientの離散化
    - グラディエントは重みそのものよりも小さいはずなので、精度を落としても良いはず
    - 重みは32bitでグラディエントは16bit/8bitなど
  - 1-bit SDG
    - 離散化の極端な例として1ビットにしたもの
    - どちら向きかだけを送る
    - 誤差拡散で1-bitでも場合によっては十分

誤差拡散

発表の概要

- 研究背景と概要

- 背景
  - データ並列機械学習
  - ネットワーク
  - SimGrid

- パラメータ交換手法

- 評価
  - 設定
  - 結果

- 関連研究

- 結論

結論

- パラメータ交換手法とネットワーク構造の関係
  - 並列シミュレータ「SIMGRID」を用いて調査

- パラメータ交換手法を比較
  - サーバ間で直接交換するほうがパラメータサーバを用いる場合よりも高速
  - 直接交換する場合、階層的なバタフライを行うと比較的プアなネットワークでも速度が低下しない
  - パラメータサーバを用いる場合、パラメータサーバの配置を工夫することで、ネットワークの影響を低減可能

今後の課題

- リアリスティックなパラメータでの評価
  - ネットワークバンド幅
  - パラメータ交換の間隔
  - パラメータデータ量
- ネットワークトポロジの影響の調査
  - サブクラスタのサイズ、階層数
  - Cyclic Banyan Network?
- さまざまな先進的手法の影響の調査
  - 非同期
  - ゴシップ通信
  - グラディエントの離散化
  - 1bit SDG

謝辞

- この成果の一部は、国立研究開発法人新エネルギー・産業技術総合開発機構 （ＮＥＤＯ）の委託業務の結果得られたものです。
- 本研究はJSPS科研費 JP16K00116の助成を受けたものです。