

【ICDE2015勉強会】

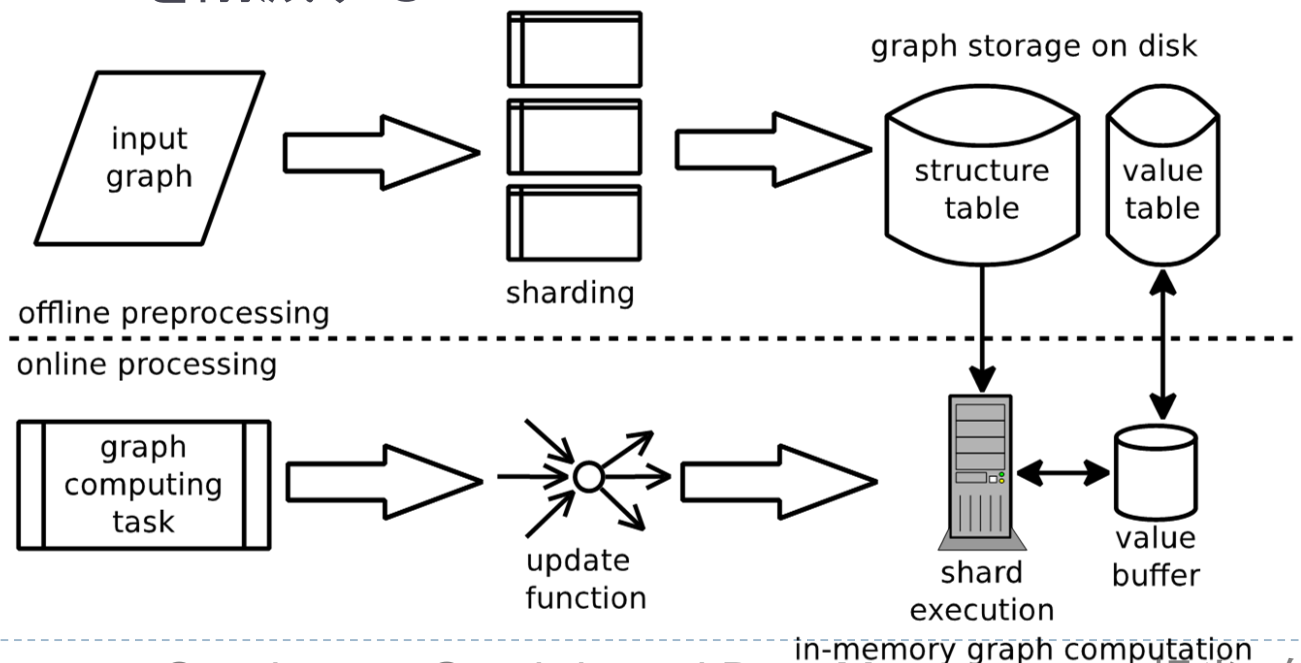
Session 25: Graph-based Data Management

担当：鬼塚 真(大阪 大学)

本資料の一部には、著者の論文・スライドのコンテンツを利用しています

論文: VENUS: Vertex-Centric Streamlined Graph Computation on a Single PC, Jiefeng Cheng, Qin Liu, Zhenguo Li, Wei Fan, John C.S. Lui, Cheng He

- ▶ 特徴: グラフをshard 分割, shard単位でストリーム処理する. Shard はエッジの到達ノードをキーに範囲分割
 - ▶ 到達ノードをキーとしてグラフを shard を構築する
 - ▶ 変更される箇所と変更されない箇所(グラフ構造)を分離してIOコストを削減する



shard の構造

- ▶ データ構造: Incoming edge の情報でshard を構築する
 - ▶ G-shard: 更新されないグラフ構造部分
 - ▶ V-shard: vertex 情報(更新可能)

Interval	$I_1=[1,4]$	$I_2=[5,8]$	$I_3=[9,12]$
G-shard	7, 9, 10 → 1 6, 10 → 2 1, 2, 6 → 3 1, 2, 6, 7, 10 → 4	6, 7, 8, 11 → 5 1, 10 → 6 3, 10, 11 → 7 3, 6, 11 → 8	2, 3, 4, 10, 11 → 9 11 → 10 4, 6 → 11 2, 3, 9, 10, 11 → 12
V-shard	$I_1 \cup \{6, 7, 9, 10\}$	$I_2 \cup \{1, 3, 10, 11\}$	$I_3 \cup \{2, 3, 4, 6\}$

- ▶ 比較: GraphChi ではvertex 周辺の incoming/outgoing edges をまとめて shard を構築するため, メモリを効率よく使えない

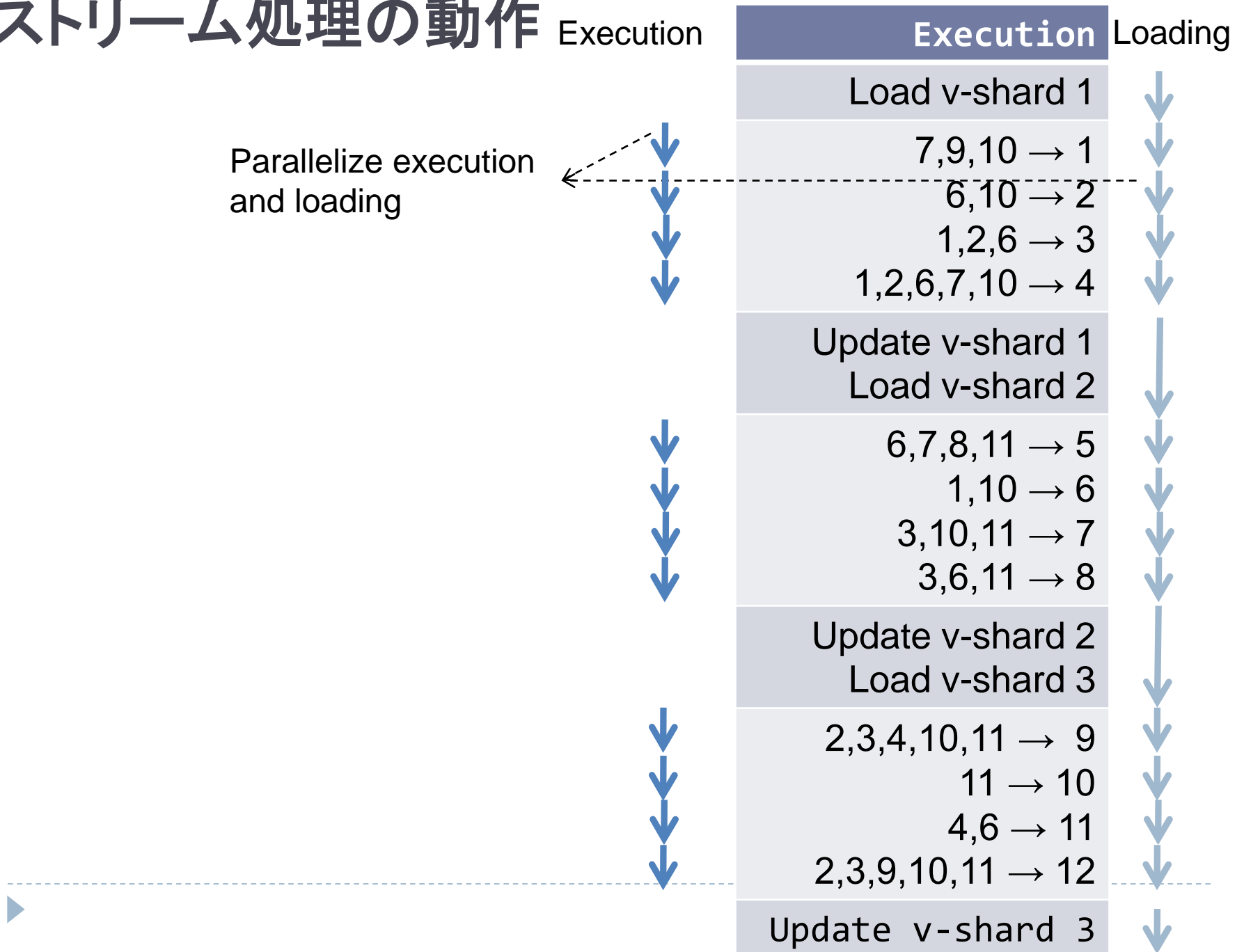
ストリーム処理の動作

Execution

Execution

Loading

Parallelize execution
and loading



Evaluation of VENUS

- ▶ **Setup: a commodity PC**
 - ▶ quad-core 3.4GHz CPU
 - ▶ 16GB RAM and 4TB hard disk
- ▶ **Main competitors:**
 - ▶ GraphChi and X-Stream
- ▶ **Applications:**
 - ▶ **PageRank**
 - ▶ **WCC**: weakly connected components
 - ▶ **CD**: community detection
 - ▶ **ALS**: alternating least square for collaborative filtering
 - ▶ Shortest path, label propagations, etc.



評価結果(参照系): pagerank計算

- ▶ IOの削減効果で性能が最大3倍程度改善
- ▶ データ: Twitter follow-graph: 41M nodes, 1.4B edges

