Lazy Snapshots: A Taxonomy and Performance Study

Zhijun Liu and Paul Sivilotti*
(and Nigamanth Sridhar)

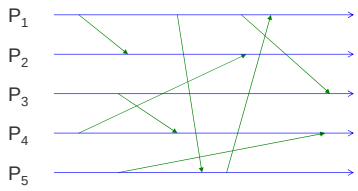
Computer Science and Engineering
The Ohio State University

*currently visiting faculty at University of Minnesota

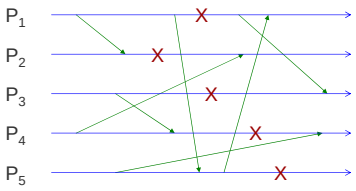
Background: Checkpointing

- Checkpoint = recording state of system
- Utility:
  - Fault recovery
  - Debugging
  - Stable property detection
- State of a distributed system consists of:
  - local state of individual processes
  - state of channels (ie messages in transit)
- Simple algorithm:
  - Every process records its local state at time t
  - Channel state?
- Challenge:
  - No shared clock

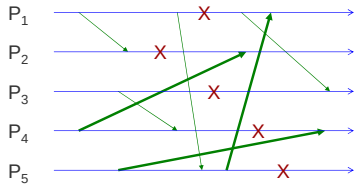
Consistent Checkpoints



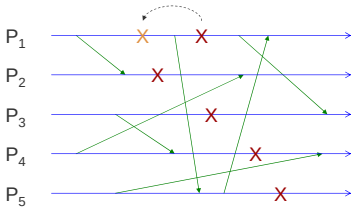
Consistent Checkpoints: Local State



Consistent Checkpoints: Channel State



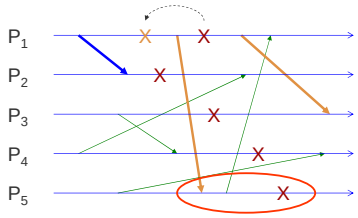
Consistent Checkpoints: Requirement



- The cut must be "input-closed":
  Every message recorded as received, must also have been recorded as sent

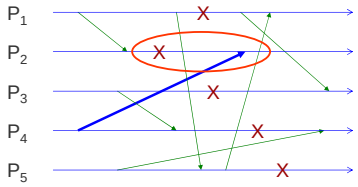
Consistency:

No *after* message should be received *before* RLS



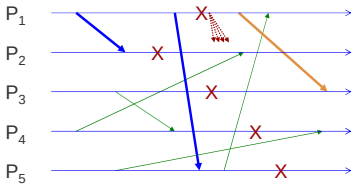
Completeness (channel state):

before messages rec'ved *after* RLS are in transit



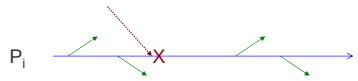
A Cooperative Snapshot Alg

- Use marker to separate *before* messages from *after* messages
  - Assumption: FIFO channels



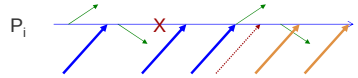
Rule 1:

Record local state *immediately* when *first* marker arrives



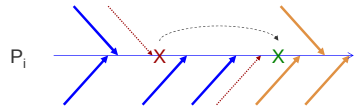
Rule 2:

Record channel state *after* RLS and *until* marker arrives



Observation: Overconstraint

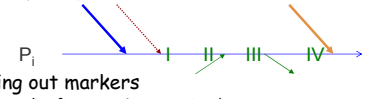
- Consistency only requires RLS before first *after* receive
  - Could be "lazy" and delay RLS



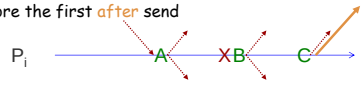
- Potential advantages:
  - Fewer messages in transit
    - less storage required to record channel state
  - Some flexibility in choosing when to RLS
    - choose less critical time
    - choose time when local state is smaller

Taxonomy of Laziness

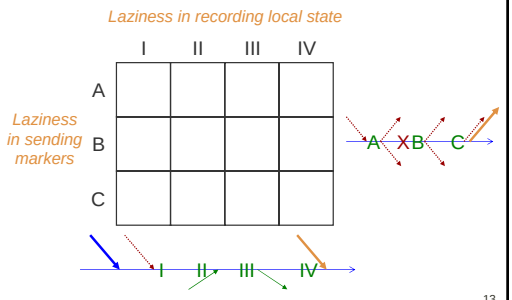
- Recording Local State
  - I. When first marker received
  - II. Before first send or receive (after rec'd marker)
  - III. Before first send or after receive (after rec'd marker)
  - IV. Before first after receive



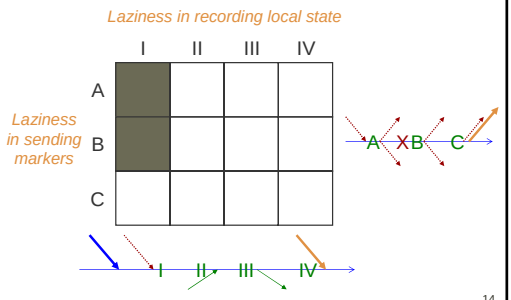
- Sending out markers
  - A. When the first marker received
  - B. When local state is recorded
  - C. Before the first after send



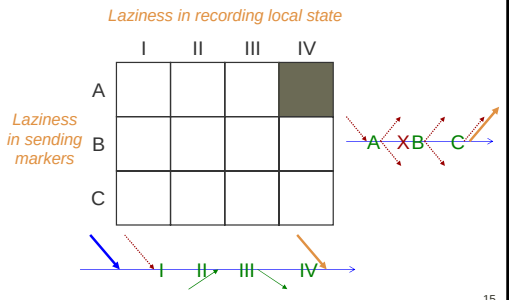
Orthogonal Dimensions?



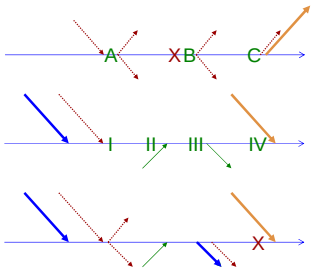
Example 1: Chandy-Lamport



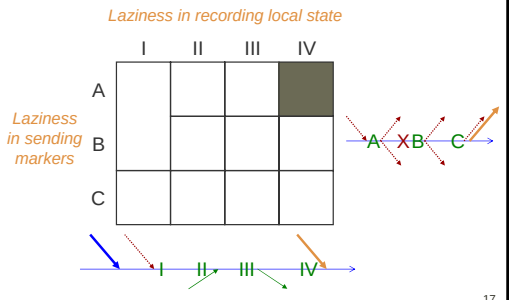
Example 2: Prescience (A-IV)



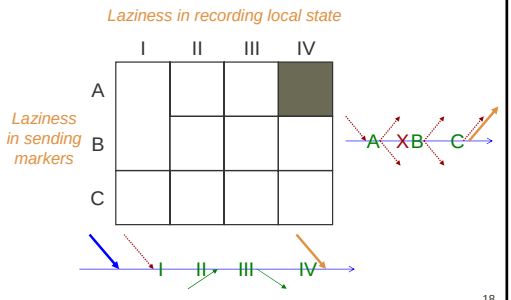
Example 2: Prescience (A-IV)



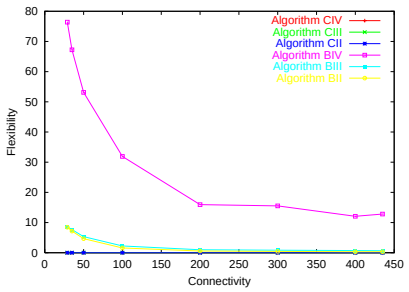
Taxonomy of Laziness



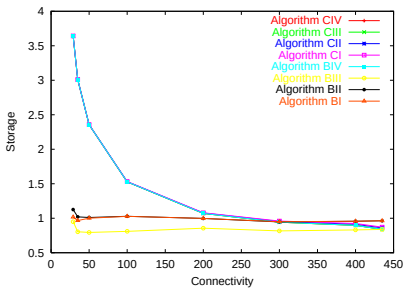
Analysis: Flexibility



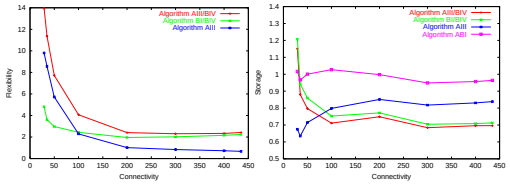
Results: Flexibility



Results: Storage



Results: Hybrid Algorithms



Conclusions

- Taxonomy with 2 dimensions
  - Laziness in recording local state
  - Laziness in propagating markers
- Dimensions are not independent
  - Chandy-Lamport: AB-I
  - Unimplementable (without prescience): A-IV
- Performance evaluation
  - Flexibility and storage complexity
- Hybrid: blending A-III / B-IV
  - Dynamic mixing based on prediction of future communication events

Lazy Snapshots: A Taxonomy and Performance Study

Zhijun Liu and Paul Sivilotti

Computer Science and Engineering
The Ohio State University

paolo@cse.ohio-state.edu