Scalable Failure Recovery for Tree-Based Overlay Networks

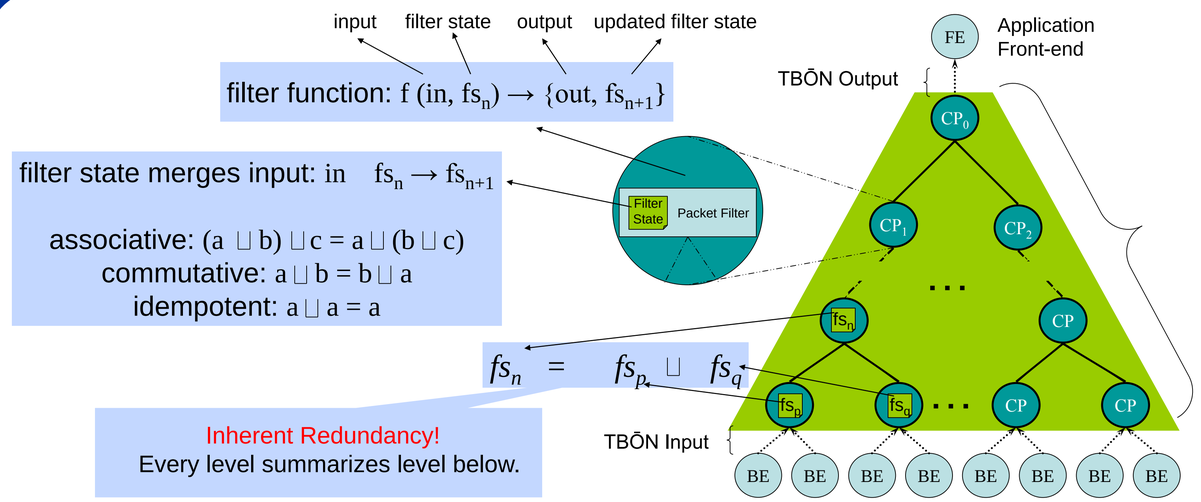
Paradyn™

WISCONSIN — UNIVERSITY OF WISCONSIN–MADISON

THE UNIVERSITY of NEW MEXICO

Tree-Based Overlay Networks

TBŌNs use a tree of communication processes with filtering capabilities to provide applications with scalable data multicast, data gather and data aggregation.



- TBŌN model applies to many computations
  - Simple: historical min, max, count, average
  - Complex: Time aligned data aggregation, graph analysis, equivalence classes
- As applications scale, failures increase:

$$MTTF \propto \frac{1}{system\ size}$$

- Large scale systems need no/low overhead failure recovery models
  - Avoid explicit replication (e.g. checkpointing, logging)
- Failure recovery must be rapid to minimize application perturbation
  - Avoid coordination & consensus protocols

Convergent Recovery

- Eventual consistency
- Different commutations and associations of input after failure cause temporary divergence
- Post-failure output converges to non-failure case

Example TBŌN `max` Computation: Failure vs. Non-failure Output

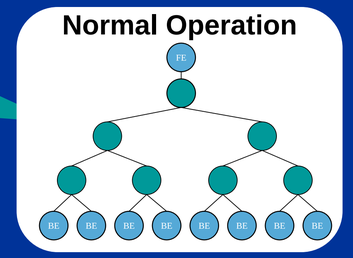
Failure may cause intermediate output divergence

	t_0	t_1	t_2	t_3	Overall Max
No Failure	7	11	27	35	35
Failure	7	8	15	35	35

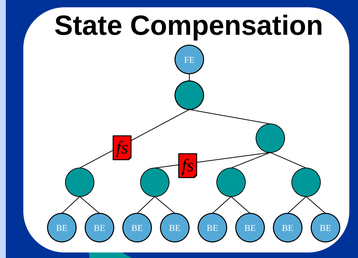
Final stream re-converges preserving all output data

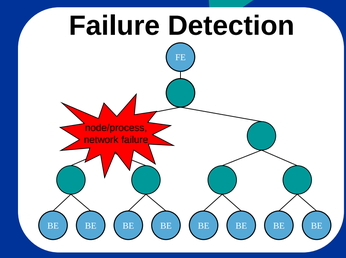
TBŌN Failure Recovery

State Compensation uses redundant information below failure zones to compensate for lost computational and communication state.



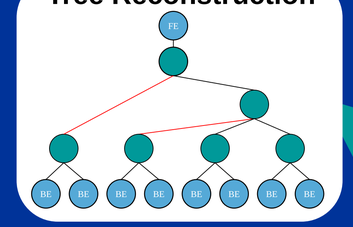
- During normal operation, state compensation does not require explicit state replication!
- Processes monitor parent & children: small group interaction.
- Child fails? Propagate notification to non-failed processes
- Parent fails? Orphaned sub-trees reconnect.





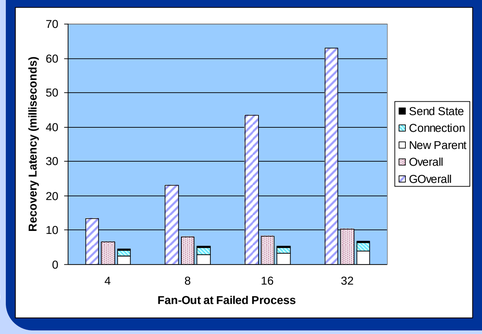
Failure Recovery Life Cycle

- For all of data flows, flush filter state as output to new parent.
- Filter state captures input history.
- Idempotence tolerates potential repeat processing of input data.



- Orphans select new parent without coordination.
- Potential parents scored by fan-out, depth, and distance from orphan.
- Avoid overloading "best" parent :
  - Scores used in weighted random sampling to rank potential parents

Early Results



Observer injects failure

Orphans report to observer after completing recovery

- Send State: Average time to propagate filter state
- Connection: Average time to connect to new parent
- New Parent: Average time to select new parent
- Overall: Average total recovery time
- GOverall: Overall recovery time viewed by external observer