Scalable Failure Recovery for Tree-Based Overlay Networks

**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:
    \[ \text{MTTF} \propto \frac{1}{\text{system size}} \]
  - Large scale systems need no/lower 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:**

<table>
<thead>
<tr>
<th>Failure vs. Non-failure Output</th>
<th>( t_0 )</th>
<th>( t_1 )</th>
<th>( t_2 )</th>
<th>( t_3 )</th>
<th>Overall Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Failure</td>
<td>7</td>
<td>11</td>
<td>27</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Failure</td>
<td>7</td>
<td>8</td>
<td>15</td>
<td>35</td>
<td>35</td>
</tr>
</tbody>
</table>

- **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.
- 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 orphans.
- Avoid overloading "best" parent: Scores used in weighted random sampling to rank potential parents

**Normal Operation**

- Observer injects failure
- Orphans report to observer after completing recovery

**Failure Detection**

- Fan-Out at Failed Process
- See [http://www.paradyn.org/mrnet](http://www.paradyn.org/mrnet)

Dorian C. Arnold, University of New Mexico