# Near-Optimal Distributed Failure Circumscription

Jake Beal MIT CSAIL



# Why is this cool?

\* Powerful Network Primitive

"Quarantine" zone for interim atomic transactions Patch-repair for routing tables Improved bounds for self-stabilizing algorithms Abstraction barrier for integrating "error mode" and "normal mode"

#### Exception handling for networks?

# Why is this different?

- \* Self-organizing distributed algorithm
- \* No infrastructure assumptions
- \* Scalable to million/billion range
- General exception framework
- \* Spatiall embedded network (less important)

### Talk Outline

- Problem: Failure Circumscription
- \* Context: Amorphous Computing
- Detecting Circumscription Locally
- \* The Algorithm
- Further Directions

### BostonNet Scenario

(Metropolitan Ad-Hoc Network)

- \* ~10^7 Nodes
- \* ~10^3 hops diameter
- \* Peer-to-peer
- \* No central control

### Network Model

- \* Embedded in Euclidean space
- \* Spatially local links (e.g. wireless)
- Perfect communication (via threshold)
- Partial synchrony (drifting clocks)
- Stopping failures (crashes)
  No partitions

### What is "Failure Circumscription"?



Connected set containing boundary of a connected or almost-connected failure.

# Long-Distance Circumscription



# "Near-Optimal Distributed"

Optimality

Minimum diameter circumscription

May be difficult to determine!

- \* Minimum spanning tree problem
- \* Big problems swamp small problemsGoal is actually smooth scaling

Distributed

Self-Organizing, Peer-to-Peer

Centralized = Vulnerable

### Context: Amorphous Computing



#### Anatomy of a Persistent Node



# PNHierarchy



.183

Level 3

### Hierarchy Requirements

- \* Uniform depth "levels"
- \* O(lg diam) levels
- \* Maximum cluster diameter  $d_i = kb^i$

 $d_i \leq b * diam$  at root

\* Neighbor relation within  $3d_i$ 

"Tight" within  $d_i$ 

# Big Idea

\* Neighbors  $\approx$  topology sketch





No Lost Neighbors

Lost Neighbors

#### "Provably Dead"



\* A set of groups D is provably dead if:

D forms a tight clique

The tight neighbors of D can be connected

No tight neighbor is still a neighbor of a group in D

#### Provable Death → Circumscription

**Theorem:** Following a failure F, let i be a level of hierarchy in which, for every member of the border clusters  $C_{Bi}$ , all of its pre-failure tight neighbors are either still neighbors or else provably dead. Then the union of neighborhoods of border clusters,  $C_{Bi} \cup N(C_{Bi})$ , contains a connected component which circumscribes the failure F.

#### Provable Death → Circumscription

**Corollary:** Following a failure F, let i be a level where some member of the border clusters  $C_{Bi}$  is no longer related to a pre-failure tight neighbor which is not provably dead. Then every cluster in  $C_{Bi}$  is related by a chain of neighbor relations to a cluster missing a non-provably dead neighbor.

# $d_i \geq diam \rightarrow Circumscription$

**Theorem:** Following a failure F, let  $d(B_F)$  be the maximum distance between any two machines in the border  $B_F$  following the failure, and  $d'(F \cup B_F)$ be the maximum distance between any two failing or border machines, before the failure. Then F is circumscribed by  $C_{B_i} \cup N(C_{B_i})$  for every level i where  $d_i \ge max(d(B_F), d'(F \cup B_F))$ 

# $d_i \geq diam \rightarrow Circumscription$

**Corollary:** Under the above conditions, any cluster contained entirely within F is provably dead following the failure.

**Corollary:** Under the above conditions, for any member of  $C_{Bi}$ , every pre-failure tight neighbor is either still a neighbor or else provably dead.

#### k-Competitive for Convex Failures

**Theorem:** For a convex failure F, let i be the minimum level for which  $d_i \ge d(B_F)$ . The diameter of the circumscription component of  $C_{Bi} \cup N(C_{Bi})$  is 11b-competitive with the diameter of an optimal circumscription (e.g. 22-competitive if  $d_i$  is powers of 2).

#### Non-Convex Failures



Don't care because it's a disaster!

# The Algorithm

- \* For each machine in the border:
  - \* Wake up level *i* neighborhood
  - \* Machines in level *i* neighborhood:
    - \* Add self to circumscription
    - Discover neighbor liveness
    - Propagate neighbor info via gossip
  - \* If some neighbor in  $B_i$  is not provably dead or alive
    - \* Increment *i* and start again

### Contributions

- Failure Circumscription Algorithm
  Competitive with optimal for convex failures
  Proportional to diameter for concave failures
- Powerful new tool for engineering failure response in distributed algorithms

Self-organizing, not centralized

Establish "Quarantine Zones" for failures

#### Further Directions

- Applications
  - Local Patch Repair for Routing Interim Atomic Data Storage
- Continuous Failure Analysis
- Partition Tolerance
- \* Distributed "Try-Catch"

### Acknowledgements

- \* Thank you for lots of help from:
  - Seth Gilbert
  - Tim Shepard
  - Nancy Lynch
  - Gerry Sussman & Hal Abelson