Similar applications: Awari + Model Checking
Solving awari

- Solved by John Romein [IEEE Computer, Oct. 2003]
  - Computed on VU DAS-2 cluster, using similar ideas as TDS
- Determined score for 889,063,398,406 positions
- Game is a draw
  
  Andy Tanenbaum:
  ```
  ``You just ruined a perfectly fine 3500 year old game``
  ```
Awari

- Based on retrograde analysis
  - Backwards analysis of search space (database)
- Partitions database, like transposition tables
  - Random distribution good load balance
- Repeatedly send results to parent nodes
  - Asynchronous, combined into bulk transfers
- Extremely communication intensive:
  - 1 Pbit of data in 51 hours (on 1 DAS-2 cluster)
**Wide-area Awari on DAS-3 grid**

- Implementation on single big cluster
  - 144 cores
  - Myrinet (MPI)
- Naïve implementation on 3 small clusters
  - 144 cores
  - Myrinet + 10G light paths (OpenMPI)
Initial insights

- Single-cluster version has high performance, despite high communication rate
  - Up to 28 Gb/s cumulative network throughput
- Naïve grid version has flow control problems
  - Faster CPUs overwhelm slower CPUs with work
  - Unrestricted job queue growth
    - Add regular global synchronizations (barriers)
Optimizations

- Scalable barrier synchronization algorithm
  - Ring algorithm has too much latency on a grid
  - Tree algorithm for barrier & termination detection
- Reduce host overhead
  - CPU overhead for MPI message handling/polling
- Optimize grain size per network (LAN vs. WAN)
  - Large messages (much combining) have lower host overhead but higher load-imbalance
- [CCGrid 2008]
Performance

- Optimizations improved grid performance by 50%
- Grid version only 15% slower than 1 big cluster
  - Despite huge amount of communication (14.8 billion messages for 48-stone database)
From Games to Model Checking

- Distributed model checking has very similar communication pattern as Awari
  - Search huge state spaces, random work distribution, bulk asynchronous transfers
- Can efficiently run DiVinE model checker on wide-area DAS-3, use up to 1 TB memory [IPDPS’09]
Wide-area performance DiVinE

![Graph showing performance comparison between different configurations: Grid, Cluster/TCP, Cluster/MX, and their respective CPU counts and optimal paths.](image)

- Grid, \((62+32+34+32) = 160\) CPUs, MAP
- Cluster/TCP, \(80 \times 2 = 160\) CPUs, MAP
- Cluster/MX, \(80 \times 2 = 160\) CPUs, MAP
Required wide-area bandwidth

![Graph showing WAN throughput vs. total number of cores for different core configurations.](image)