The Shape of Failure

Taliver Heath, Richard Martin
and Thu Nguyen

Rutgers University
Department of Computer Science

EASY Workshop
July 2001

Goal and Motivation

- Characterize workstation availability
- Scalable Internet Services
  - built from clusters for scalability and fault isolation
  - but components not designed for availability
- Current Availability methods ad-hoc
  - over-engineer and hope for the best
  - restless sleep next to pagers

Design Approach

- Decompose system into components
- Characterize fault behavior of each component in isolation
- Design system so desired overall failure rate tolerates failure rates of components
- This work: whole workstation is a component

Approximating the TTF

- Ideal: distribution of Time to Failure (TTF) of workstation
- Approximate “failure” with reboot
- TTF ≈ TTB

Methodolgy

- Collect system *last* logs
- Observe reboot times
- Collect length of time between boots (TTB)
- Fit observed data to multiple distributions to see which is most representative

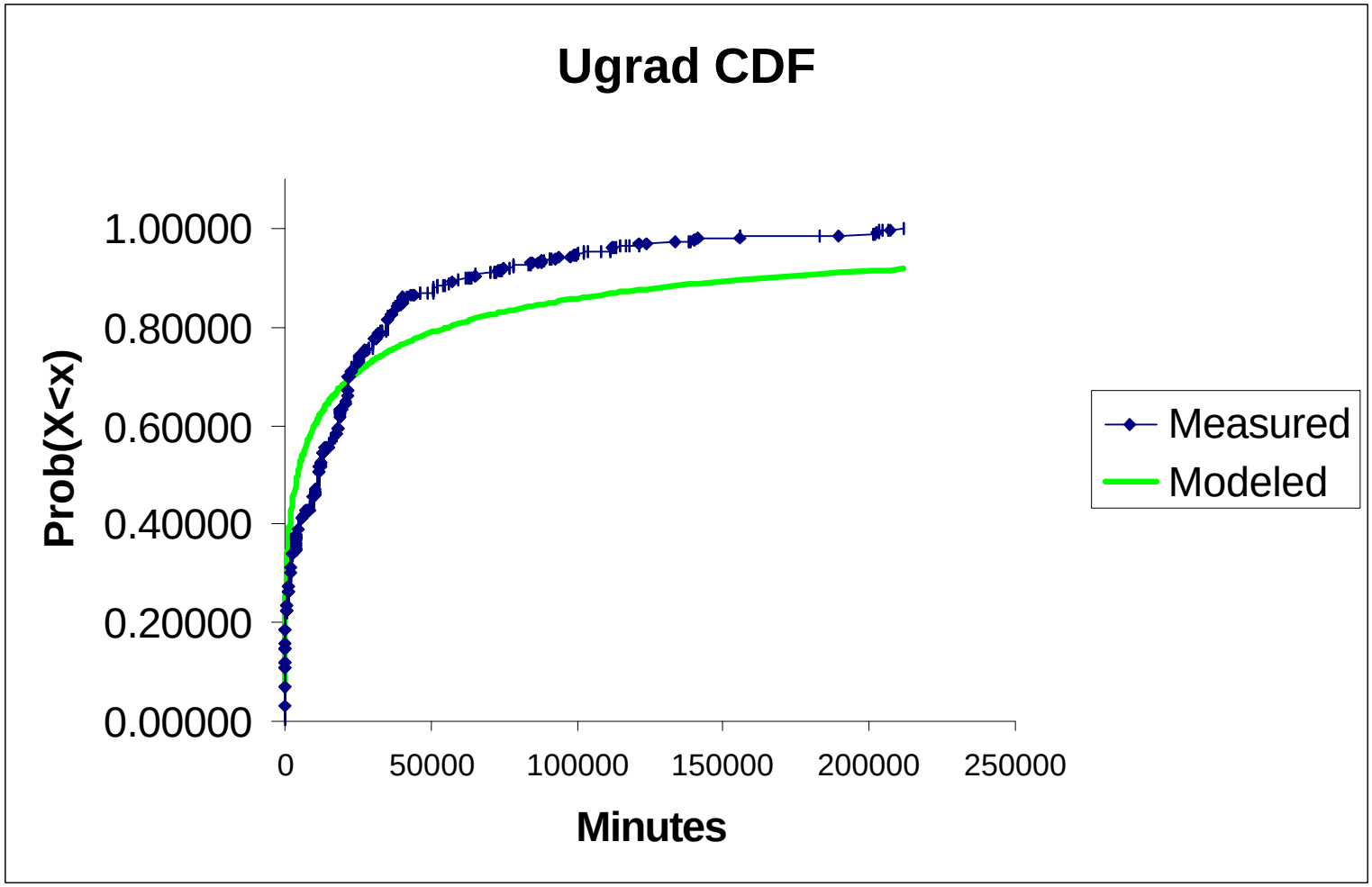
Observed Systems

- Undergrad cluster
  - 20 Ultra1 ' s open to juniors+seniors, 1 admin
- Machine room cluster
  - 17 Ultra1 ' s,2 sparc20 ' s, operator access only, 3 admins
- Industrial cluster
  - 8 Netra ' s, 9 e450 ' s, 21 Ultra ' s 1 ' s, operator access only, 1 admin

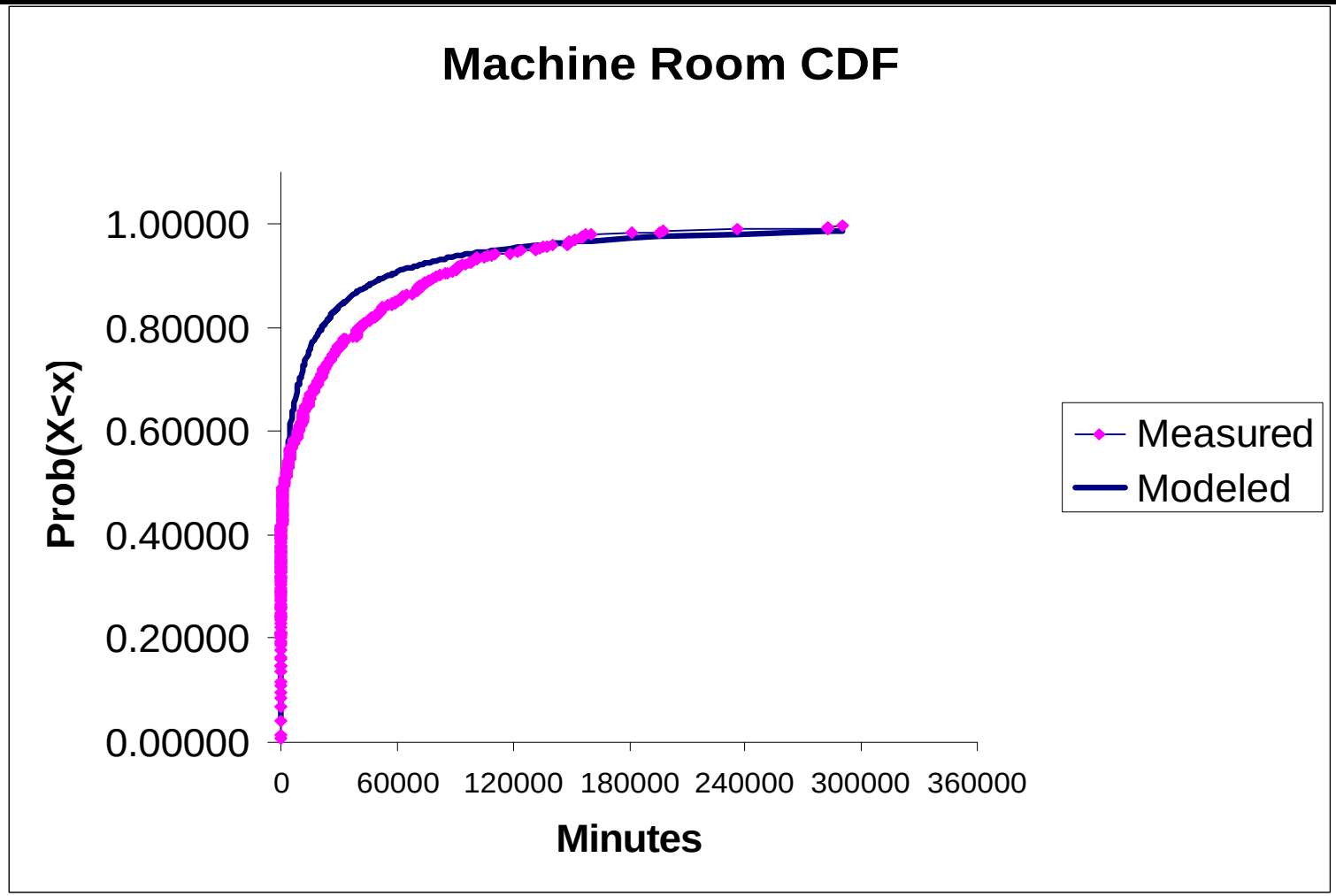
Matching to a distribution

- Maximum Likelihood Estimates to approximate the distribution
- Least squares fit to a quantile-quantile plot of data points to the distributions:
  - Exponential, Weibull, Pareto, Rayleigh
- Best match is a Weibull distribution

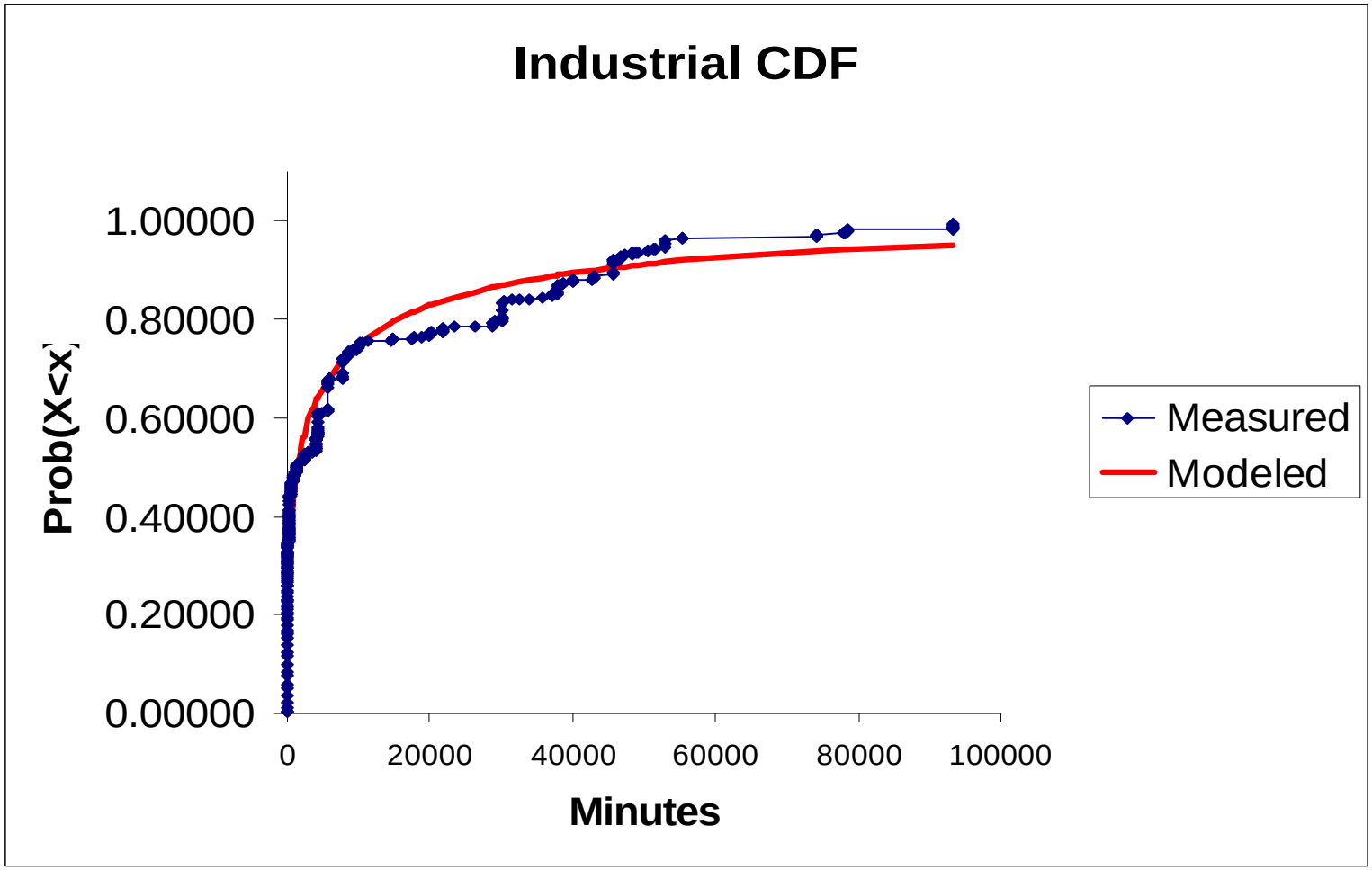
Measured vs. Modeled: ugrad



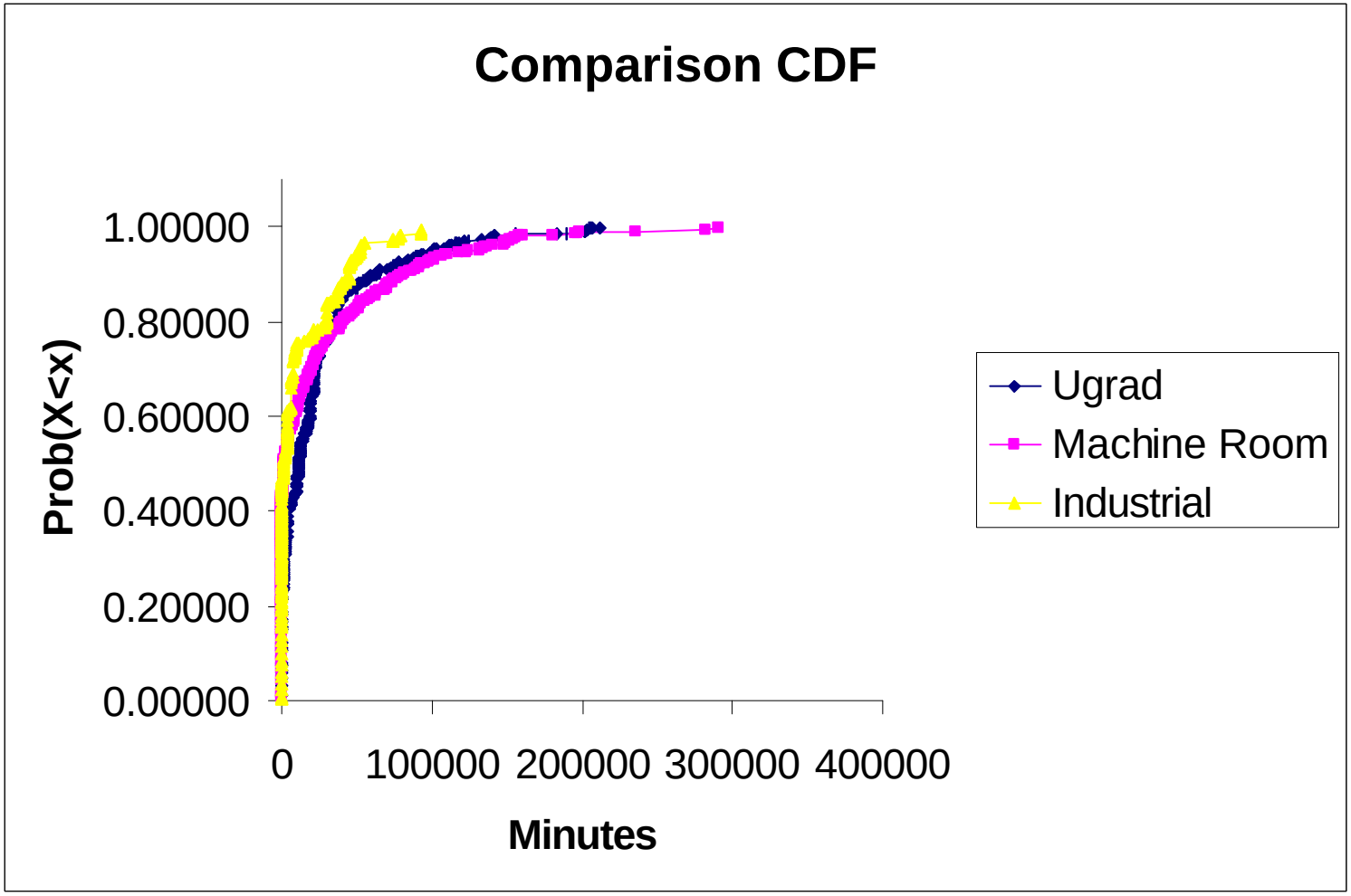
Measured vs. Molded: machine room



Measured vs. Modeled: Industrial



Side by side comparison



Results

- Workstations that have been up longer are more likely to stay up than those recently rebooted
- Weibull shape <1 mean systems not memoryless
- Similar results across all 3 clusters
  - timescales different, but shape of curves the same

Implications

- OS rejuvenation?
  - is effect large enough to observe?
- Useful lifetime < bathtub model?
  - Is a 3 year useful life < decay area?
  - All systems stay in the "flat-region"?
- Load balancing?
- Not clean when restarted?
- Upgrades

Limitations

- TTB only approximates TTF
  - e.g. a disk error may be a “failure” not captured
  - downtime not measured
- Many factors aggregated
  - difficult to determine problematic sub-component
- Independence assumption
  - model assumes independent experiments

Future Work

- Independence assumption
  - Conditional probability

    I.e. if A reboots, is B more likely to reboot soon?
- Event loggers (measurability)
  - Are reboots correlated with load?
  - What are the first-order factors?
- More/longer industrial data
- Diversification and comparison of systems
  - Same models apply to windows, linux?

More Info

www.panic-lab.rutgers.edu