Reliability Engineering Matters

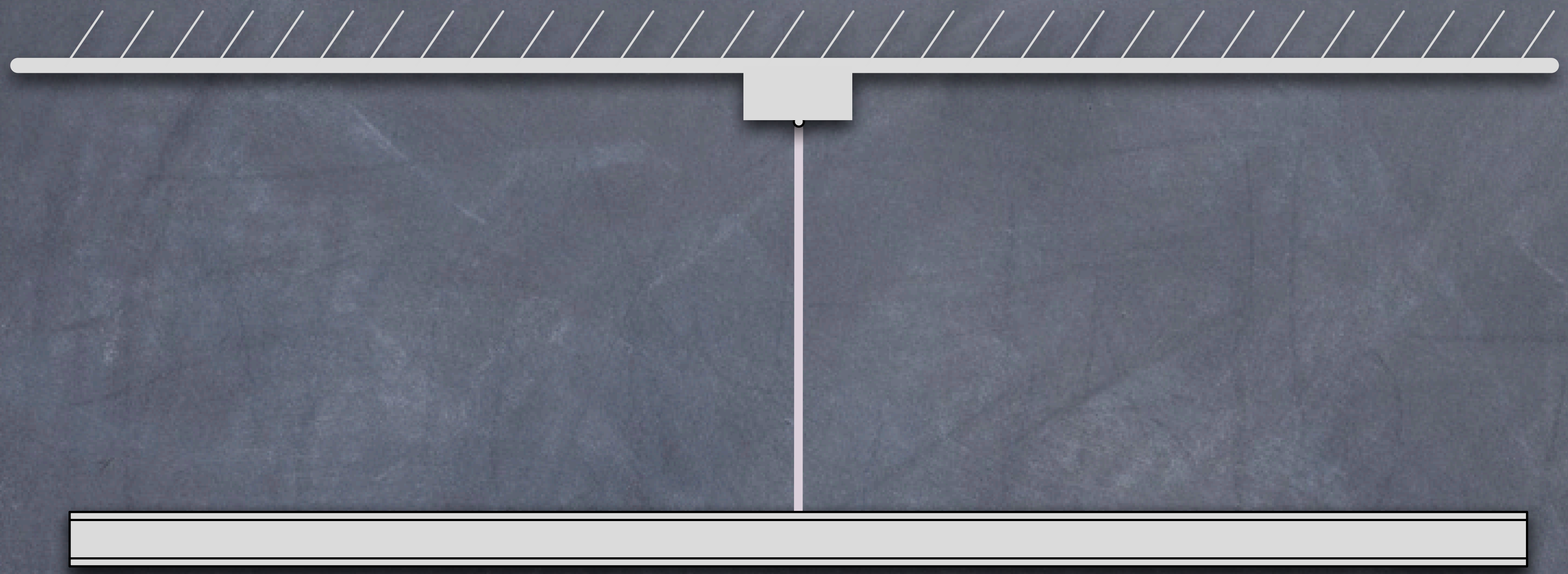
(Except when it doesn't)

Michael T. Nygard
Relevance, Inc.

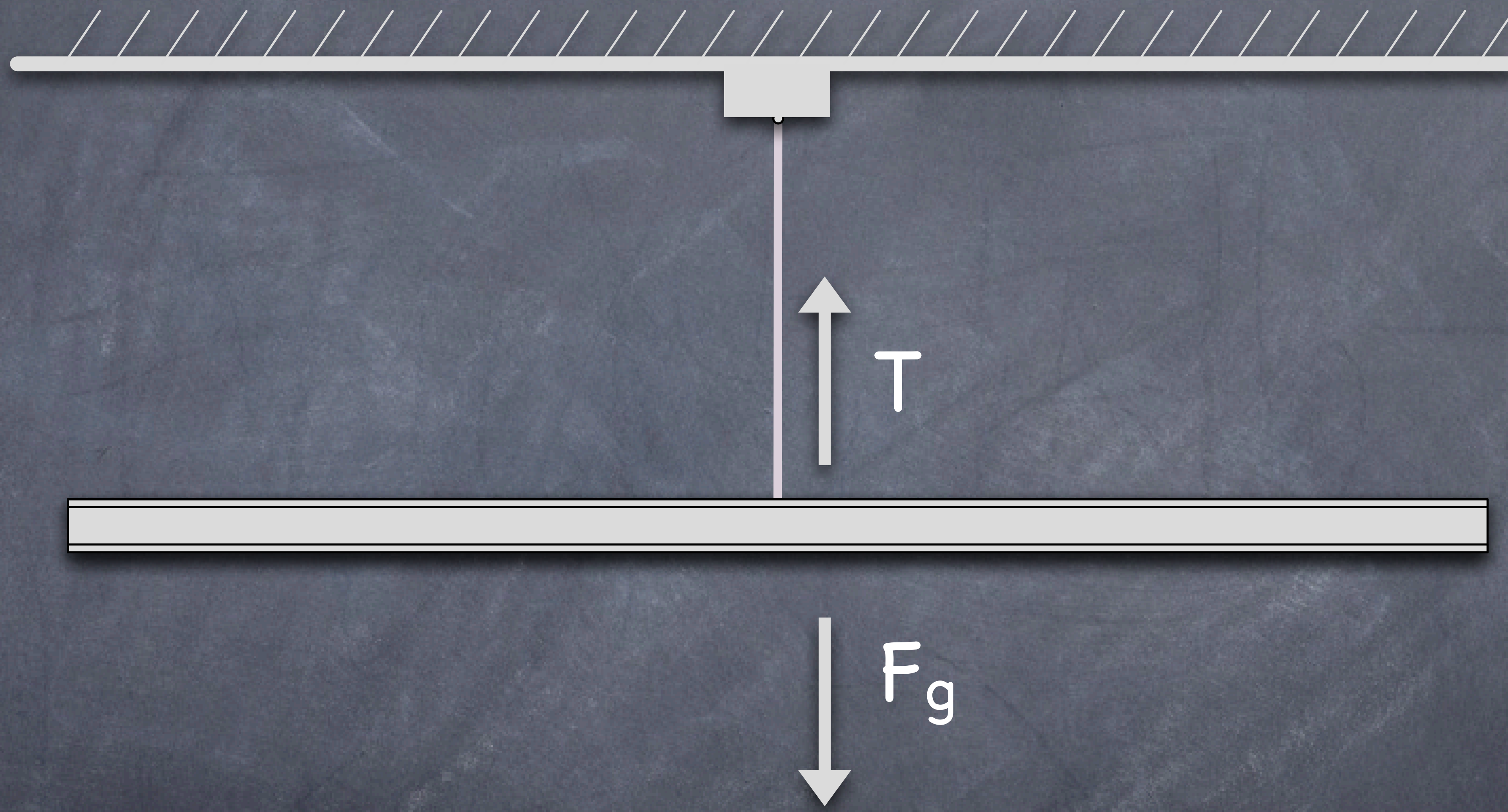
michael.nygard@thinkrelevance.com @mtnygard

Allergy Warning

This presentation contains math and math-related byproducts.

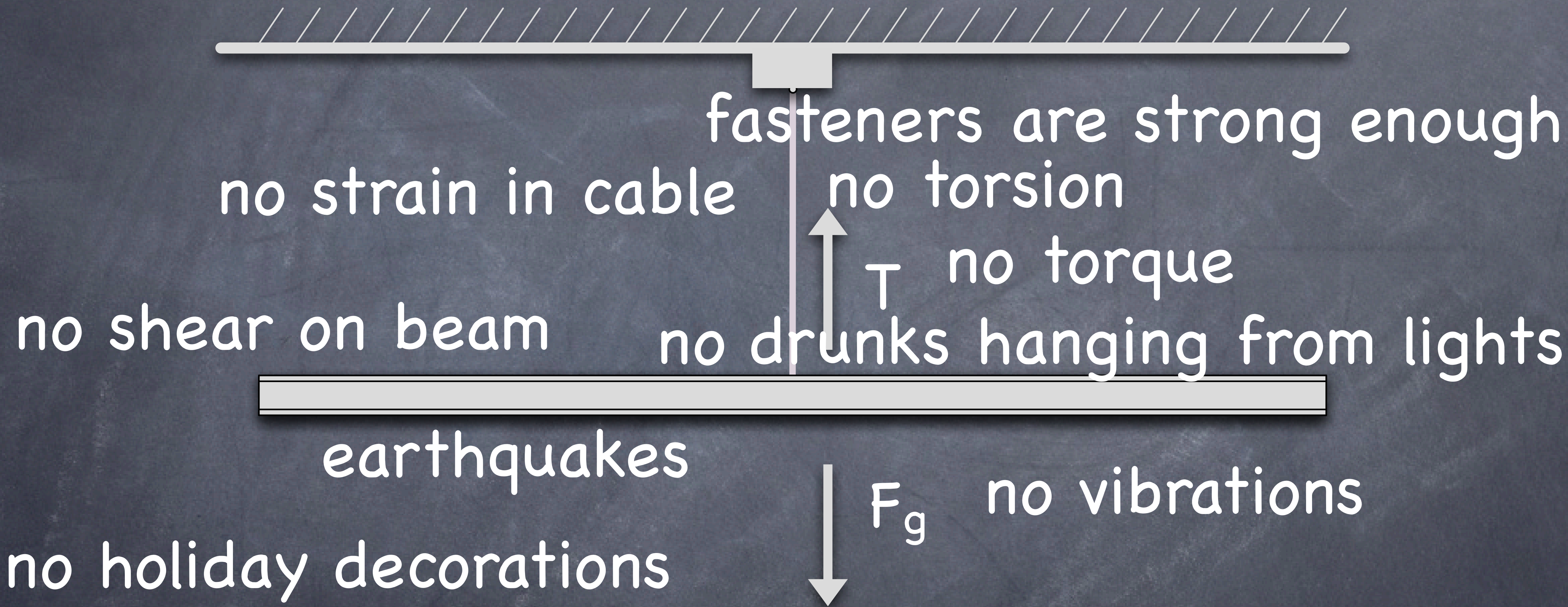


Statics & Mechanics



If $T=F_g$, everyone lives

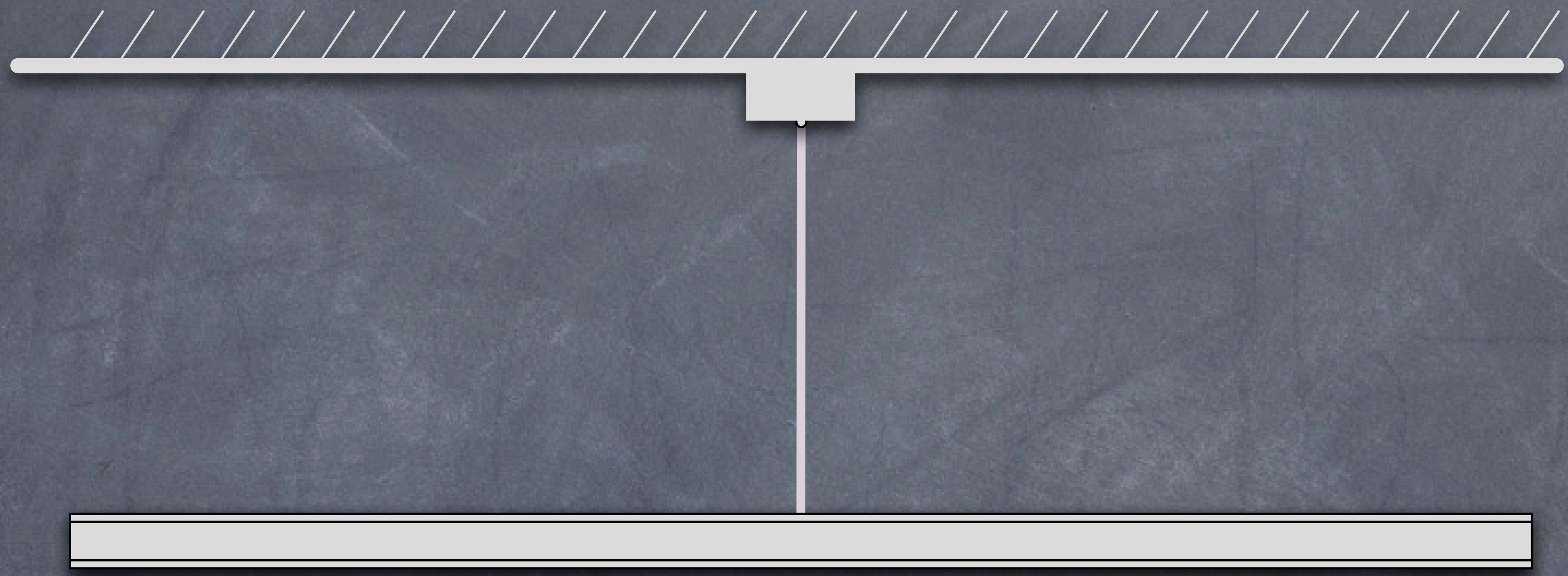
Statics & Mechanics



If $T=F_g$, everyone lives

no differential expansion
due to heat of light bulbs

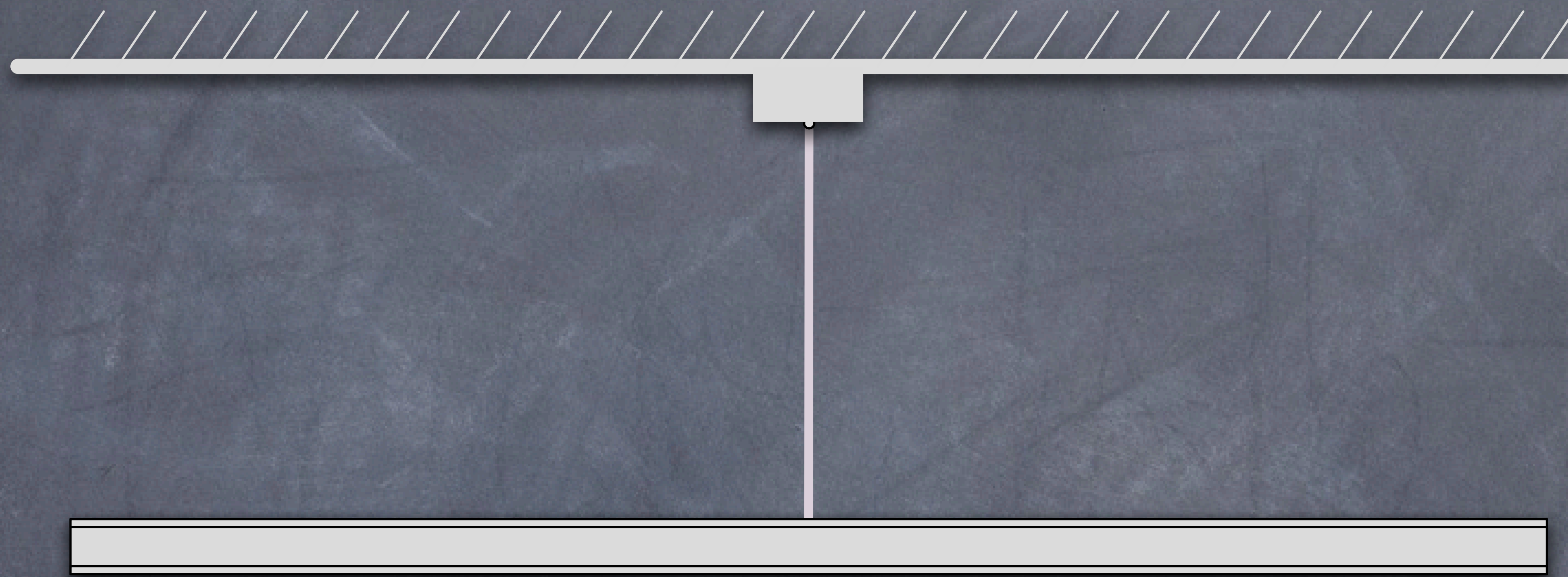
Mathematics



It didn't fall down yesterday.
Today is like yesterday.
Therefore, by the method of induction...

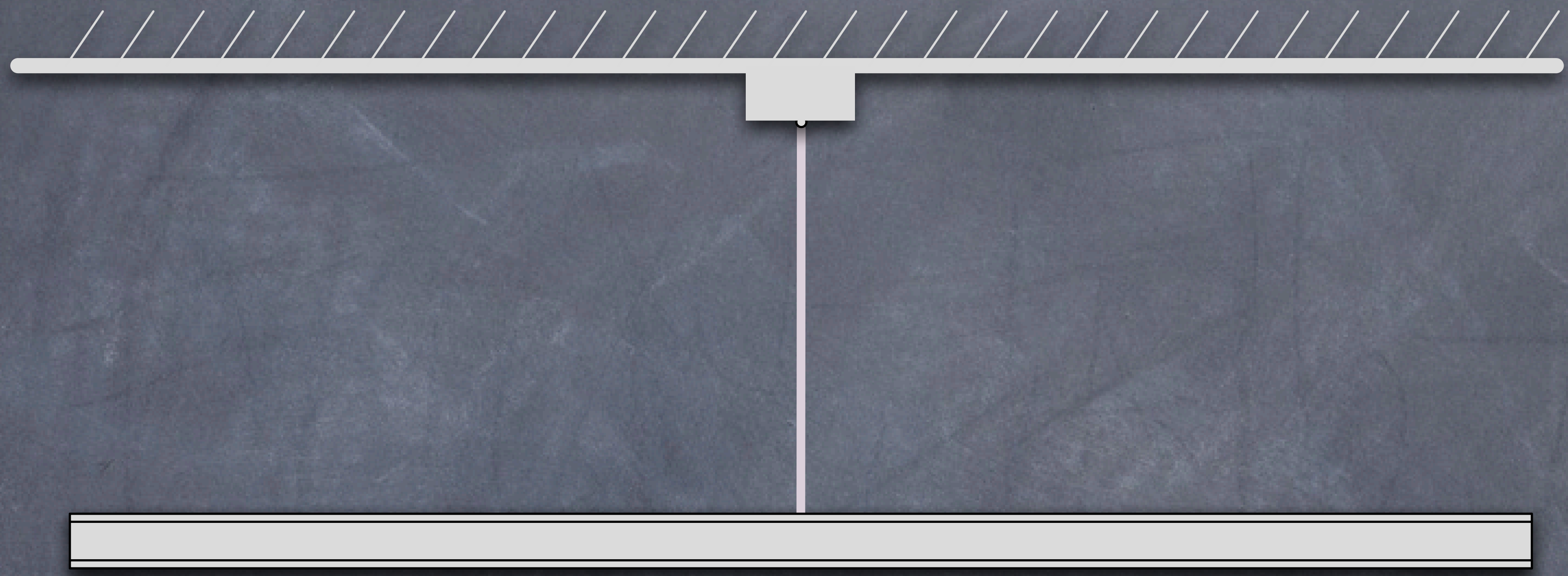
Everything breaks.
The question is when.

Reliability



How likely is it to fall down while I’m at the bar?

Reliability



T = time of failure

t = now

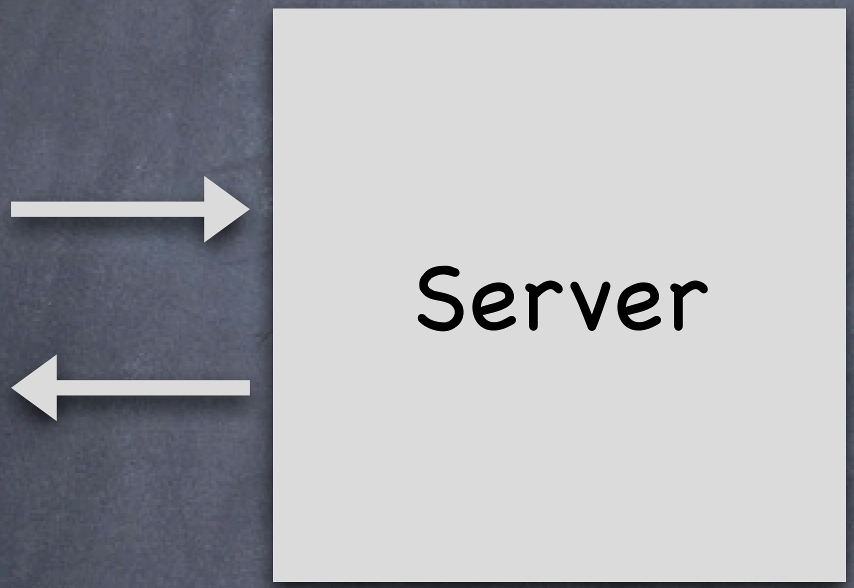
Probability of failure is $F(t) = P(\mathbf{T} \leq t)$

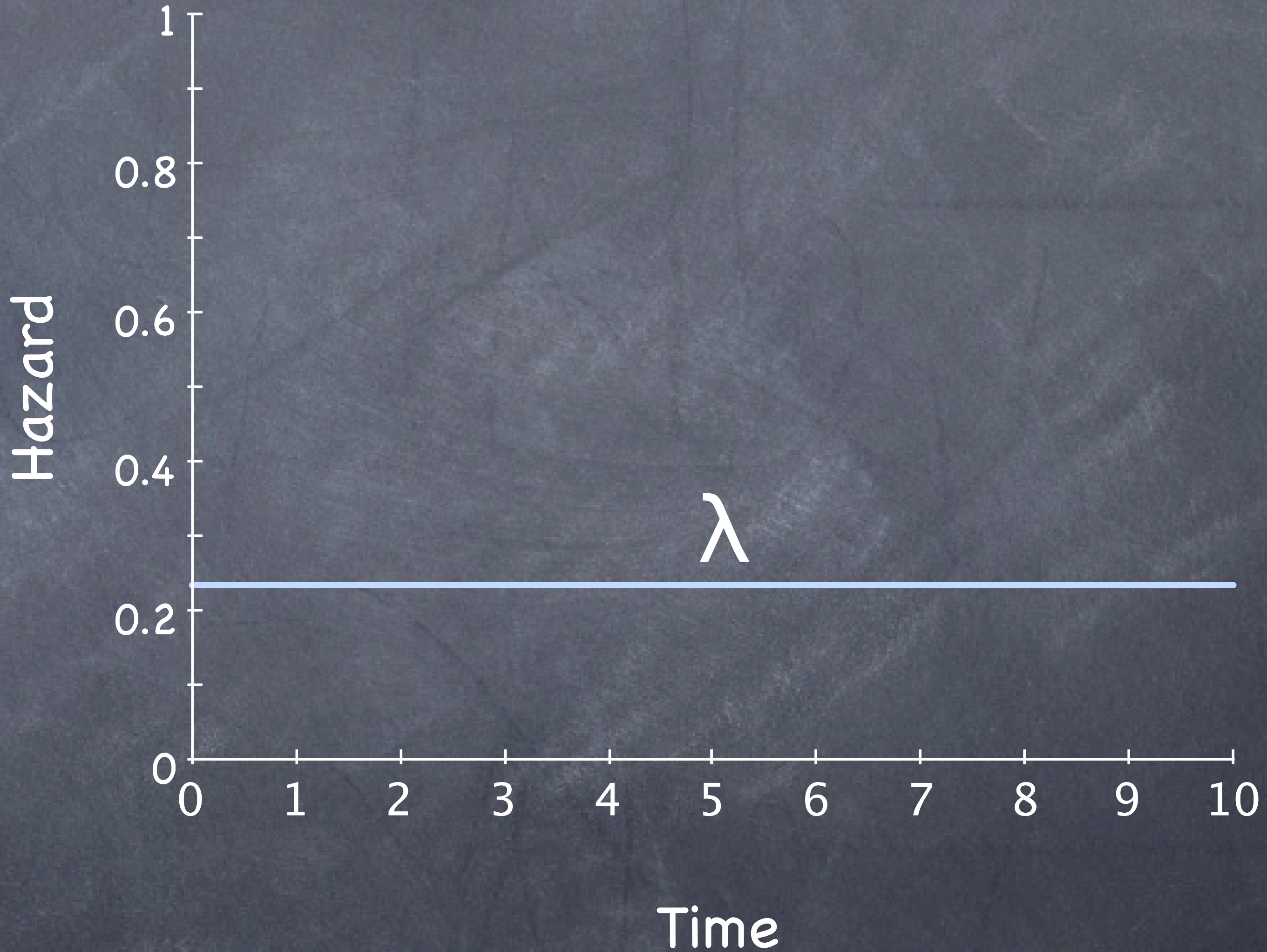
Reliability is defined as $R(t) = 1 - F(t)$

Reliability

Reliability	Did it break yet?	$R(t) = 1 - F(t)$
Failure density	Will it break soon?	$f(t) = \frac{dF(t)}{dt}$
Hazard Rate	Will it break this instant?	$z(t) = \frac{f(t)}{R(t)}$

Hazard Functions

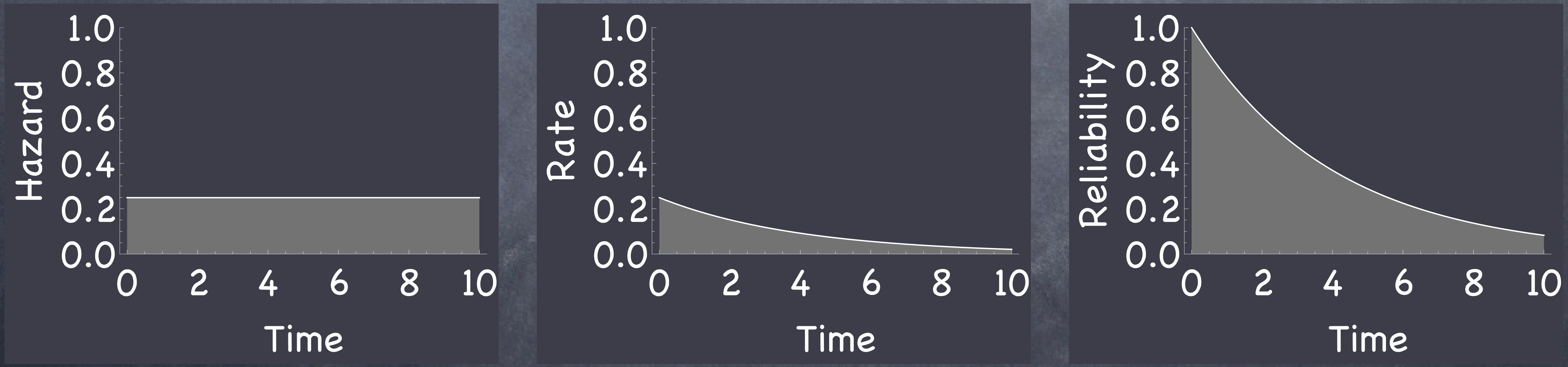




Constant Hazard

Reliability under constant hazard

Hazard	Failure Rate	Reliability
$z(t) = \lambda$	$f(t) = \lambda e^{-\lambda t}$	$R(t) = e^{-\lambda t}$



Using λ=0.25

Odds of 1 machine surviving 1 year	78%
Odds of 1 machine surviving 5 years	29%
Out of 100 machines, how many surviving after 5 years?	29

Constant hazard is what you assume when you have no actual information.

A problem has been detected and windows has been shut down to prevent damage to your computer.

If this is the first time you've seen this Stop error screen, restart your computer. If this screen appears again, follow these steps:

Check to be sure you have adequate disk space. If a driver is identified in the Stop message, disable the driver or check with the manufacturer for driver updates. Try changing video adapters.

Check with your hardware vendor for any BIOS updates. Disable BIOS memory options such as caching or shadowing. If you need to use Safe Mode to remove or disable components, restart your computer, press F8 to select Advanced Startup Options, and then select Safe Mode.

Technical information:

*** STOP: 0x0000007E (0xC0000005,0xF88FF190,0x0xF8975BA0,0xF89758A0)

*** EPUSBDSK.sys - Address F88FF190 base at FF88FE000, datestamp 3b9f3248

Beginning dump of physical memory

You need to restart your computer. Hold down the Power button for several seconds or press the Restart button.

Veuillez redémarrer votre ordinateur. Maintenez la touche de démarrage enfoncée pendant plusieurs secondes ou bien appuyez sur le bouton de réinitialisation.

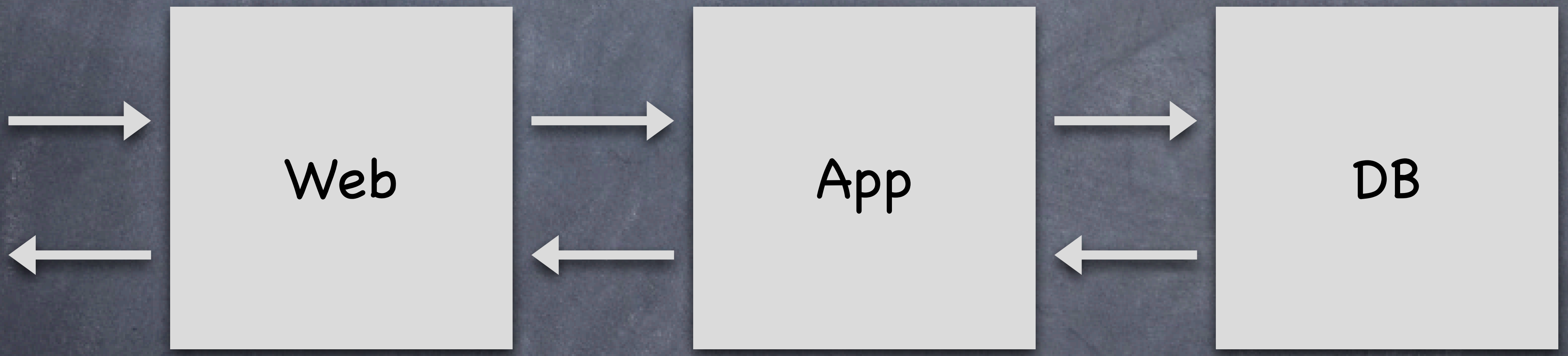
Sie müssen Ihren Computer neu starten. Halten Sie dazu die Einschalttaste einige Sekunden gedrückt oder drücken Sie die Neustart-Taste.

コンピュータを再起動する必要があります。パワーボタンを数秒間押し続けるか、リセットボタンを押してください。

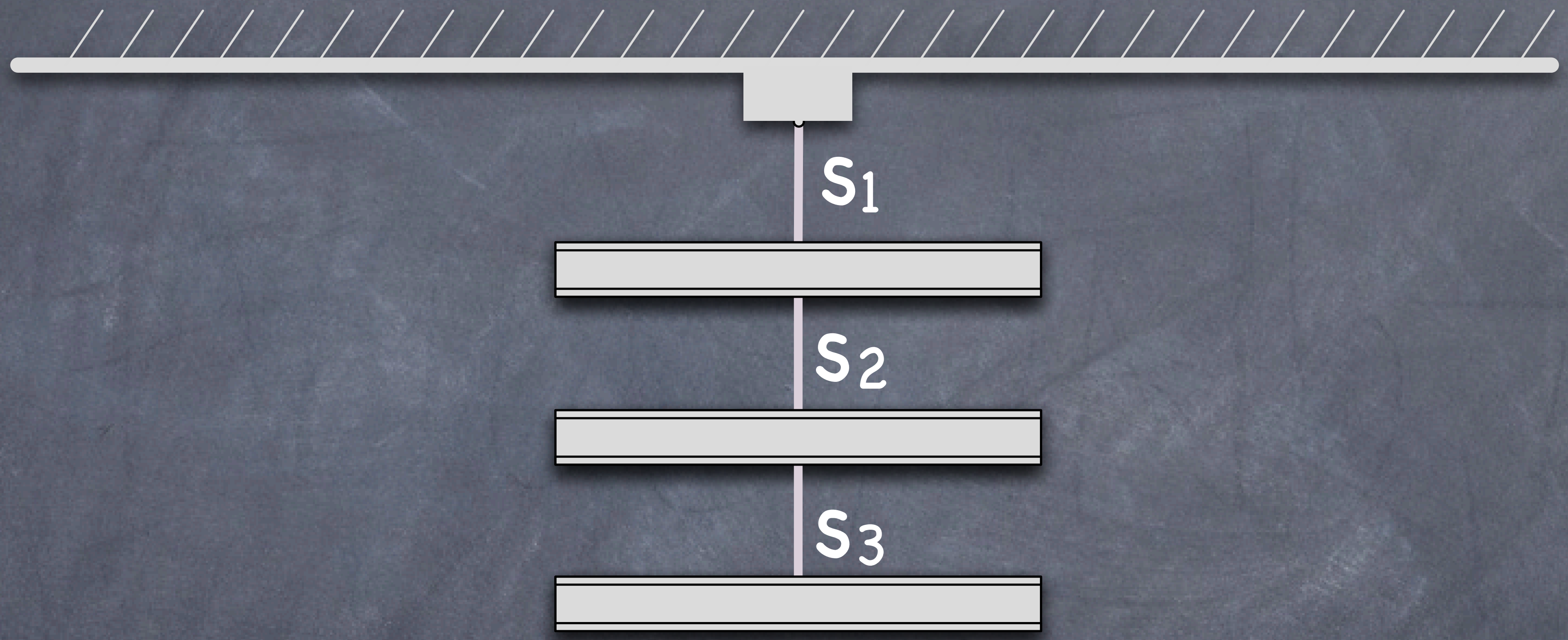
Constant hazard is what you assume when you have no actual information.

Multiple Servers

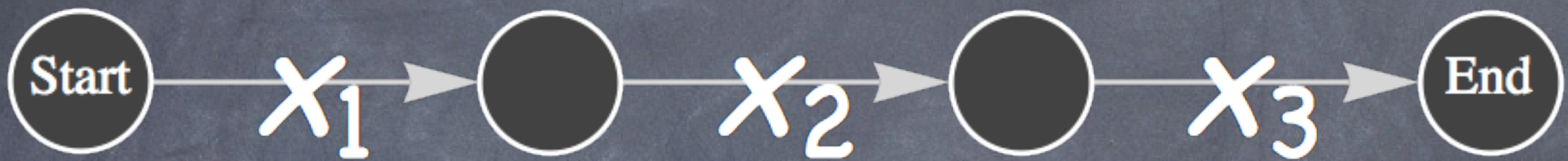
Single Strand



Single Strand



Reliability Graph-Series



Nodes are system states
Arcs are events

Probability that one unit S_n survives is $P(x_n)$
Probability of system success: $R(t) = P(x_1 x_2 x_3)$

Avoid This Fallacy

$$P(x_1 x_2 x_3) \neq P(x_1)P(x_2)P(x_3)$$

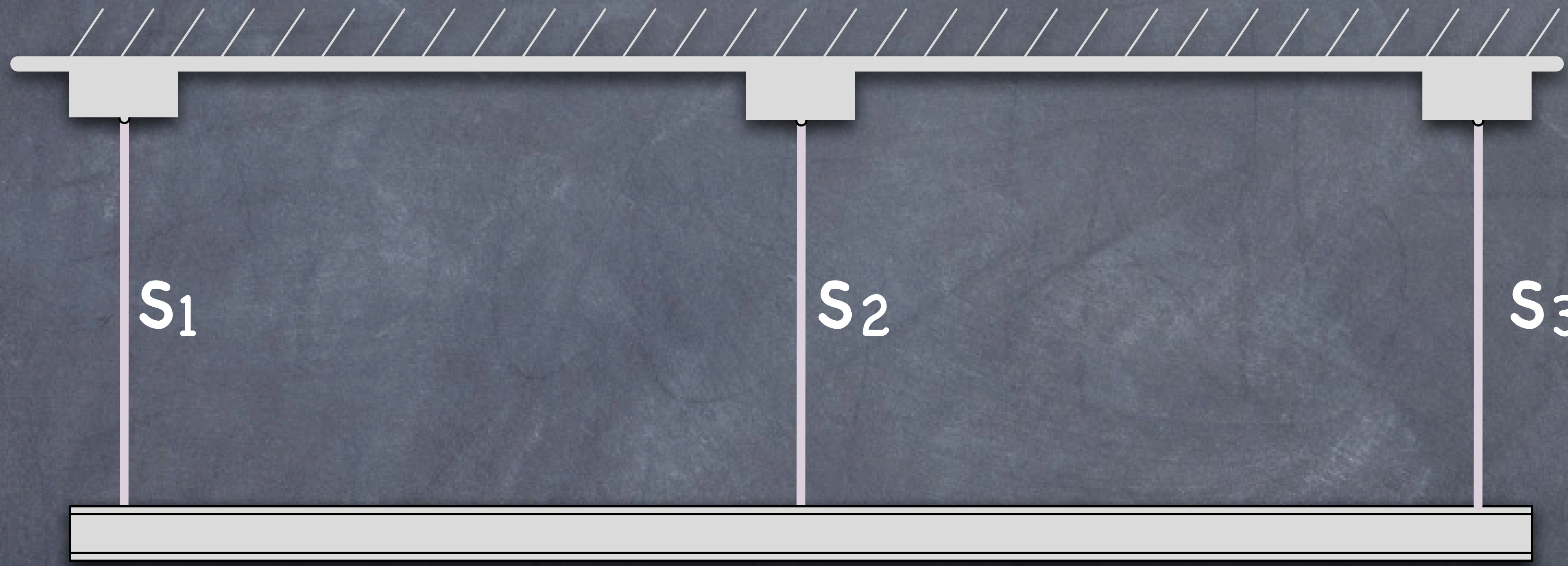
That assumes perfect independence.

$$P(x_1 x_2 x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1 x_2)$$

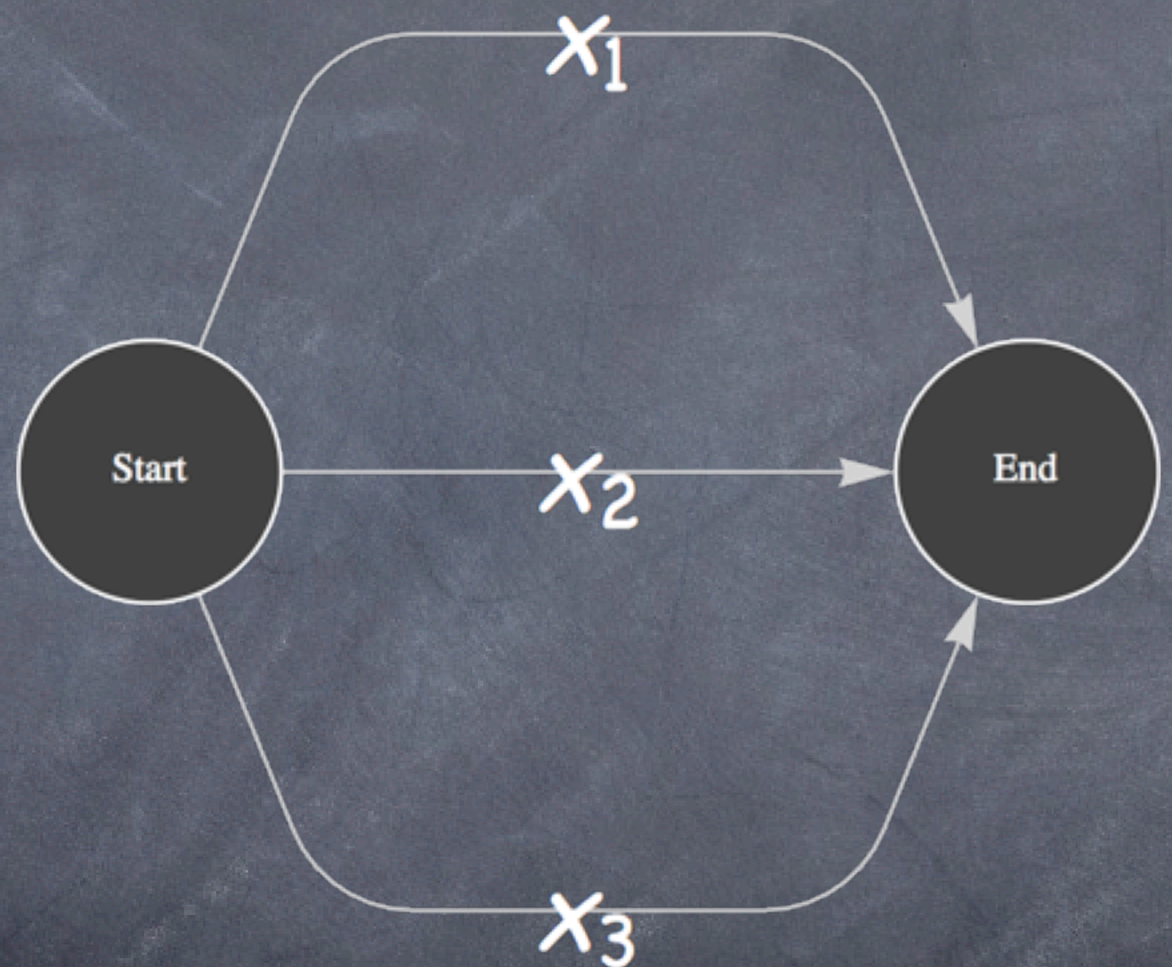
Independent	Failure of one unit does not make another unit more likely to fail.	$P(x_2\|x_1) = P(x_2)$
Correlated	Failure of one unit makes another unit more likely to fail.	$P(x_2\|x_1) > P(x_2)$
Common Mode	Something else makes both units likely to fail.	Oops, we left something out of the model.

Redundant Front End

Back to the Bar



Reliability Graph–Parallel



Probability of system success:

$$R(t) = 1-P(\overline{x}_1\overline{x}_2\overline{x}_3)$$

More #*$&#
Intersection Terms

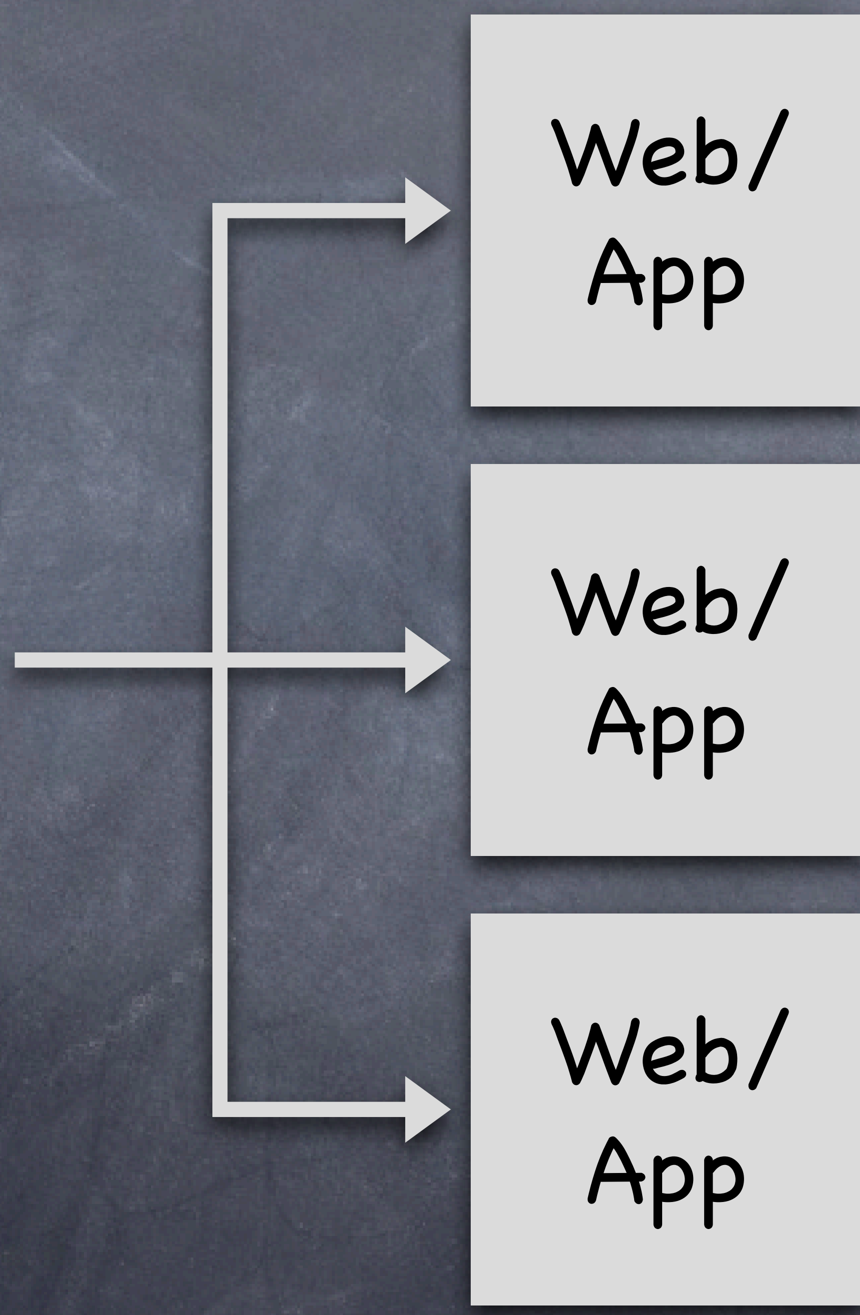
$$R(t) = 1 - P(\overline{x}_1\overline{x}_2\overline{x}_3)$$

$$R(t) = 1 - P(\overline{x}_1)P(\overline{x}_2|\overline{x}_1)P(\overline{x}_3|\overline{x}_1\overline{x}_2)$$

only if independent

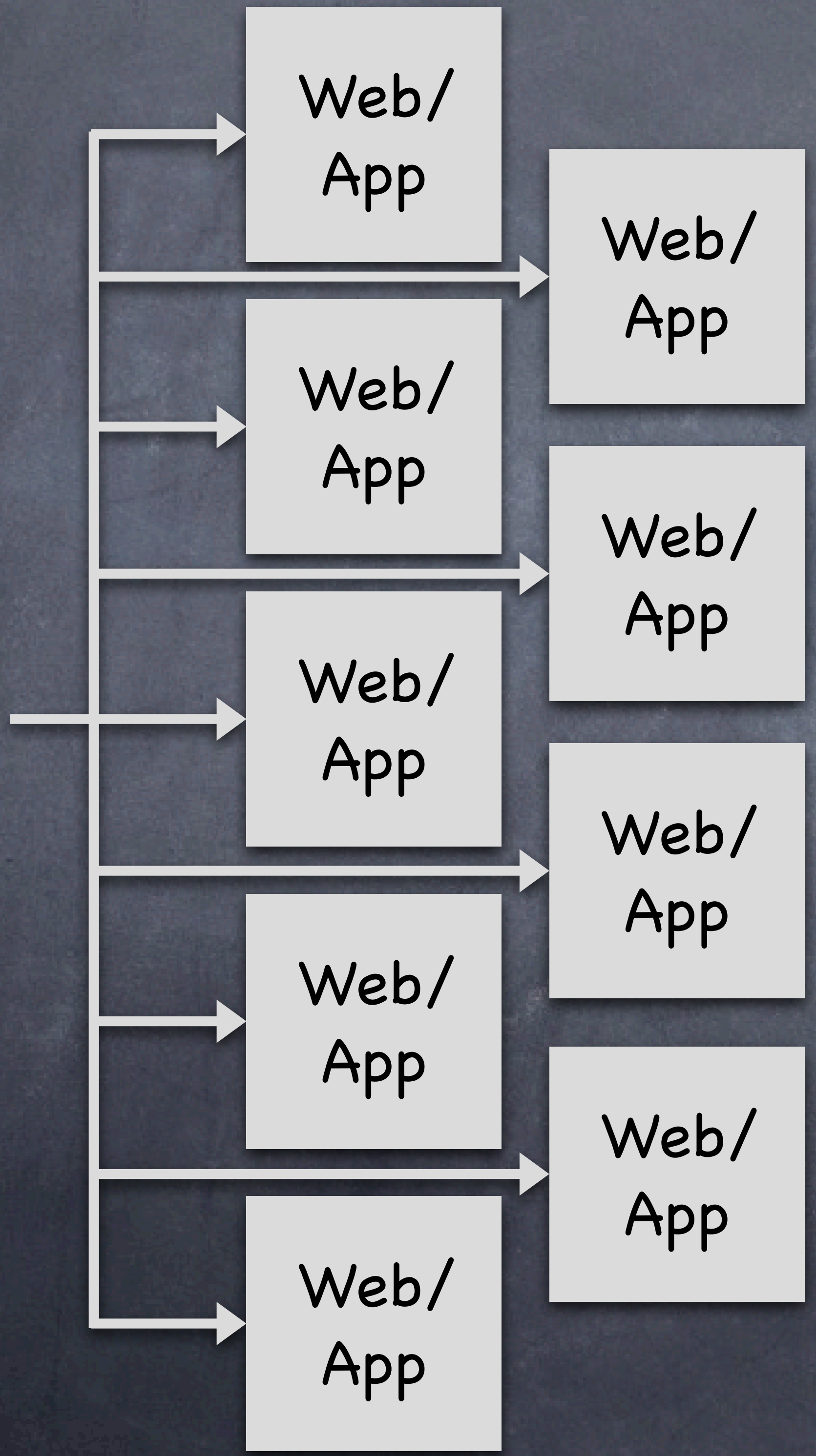
$$R(t) = 1 - P(\overline{x}_1)P(\overline{x}_2)P(\overline{x}_3)$$

Are these independent?



No!

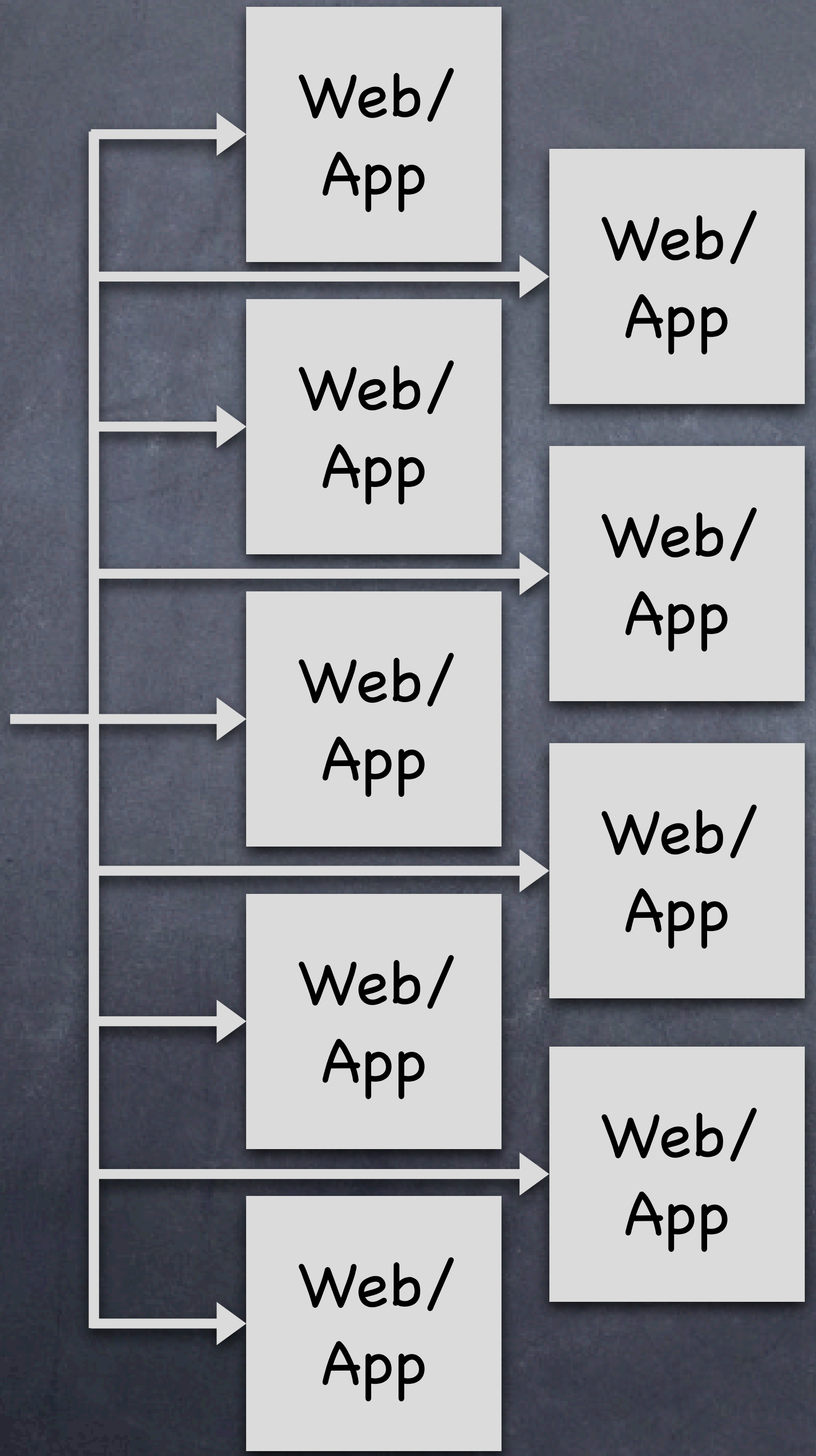
More Realism→Complexity



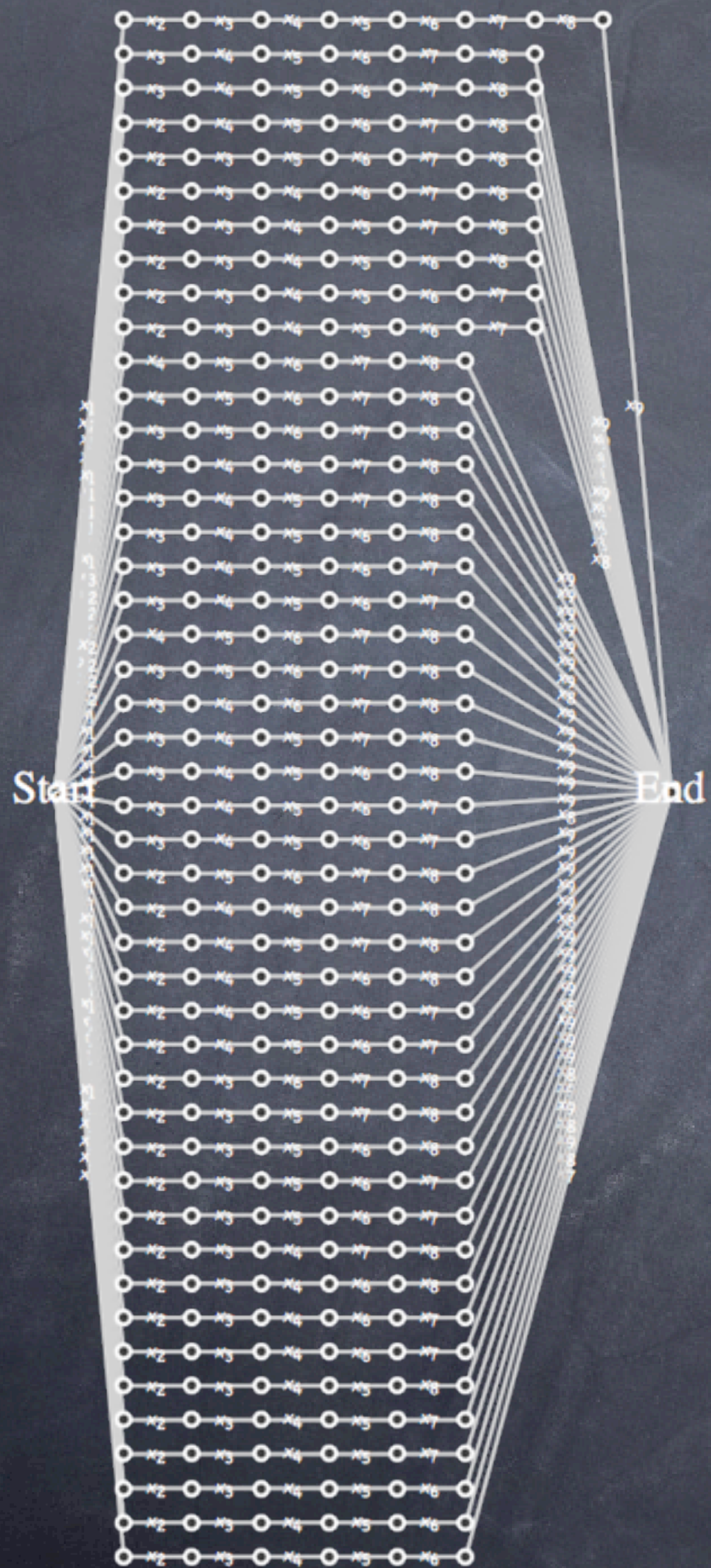
Could this system really survive on just one server?

If so, it's overbuilt.

r of k



Assume 7 of 9 required.



r of k

Assume 7 of 9 required.

Further assume:
- Identical
- Independent

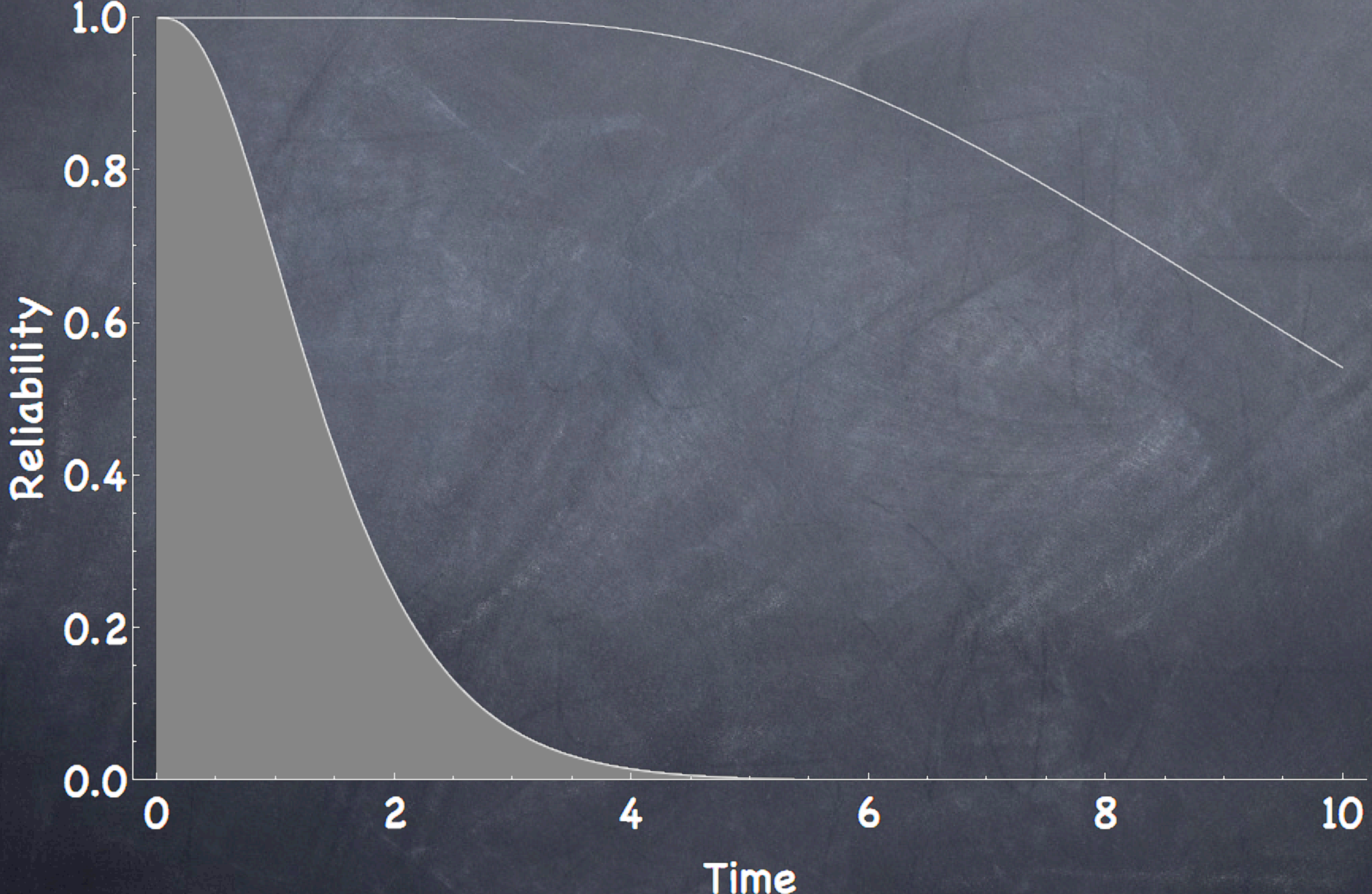
r of k

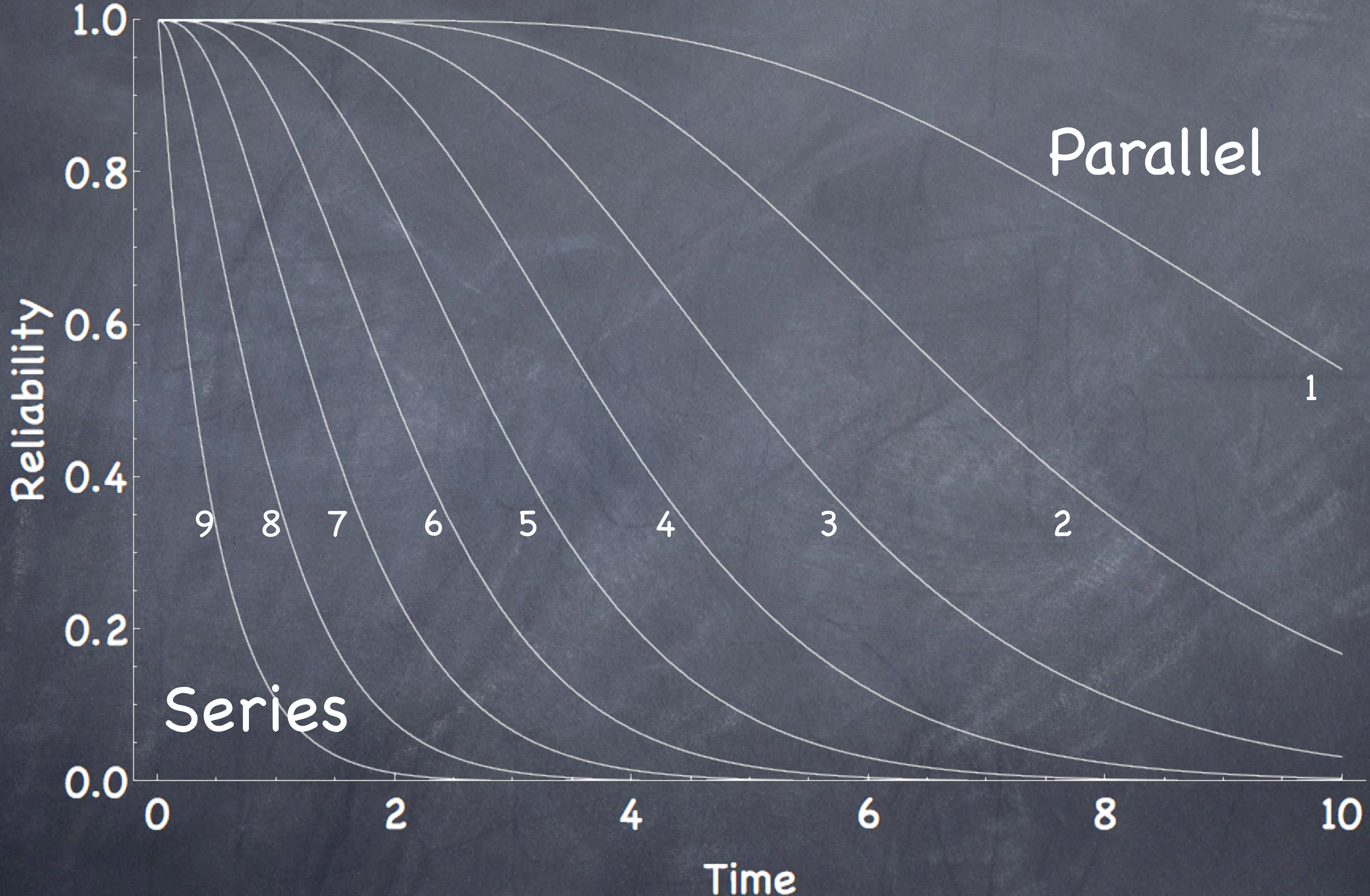
Assume 7 of 9 required.

$$R(t) = \sum_{k=r}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

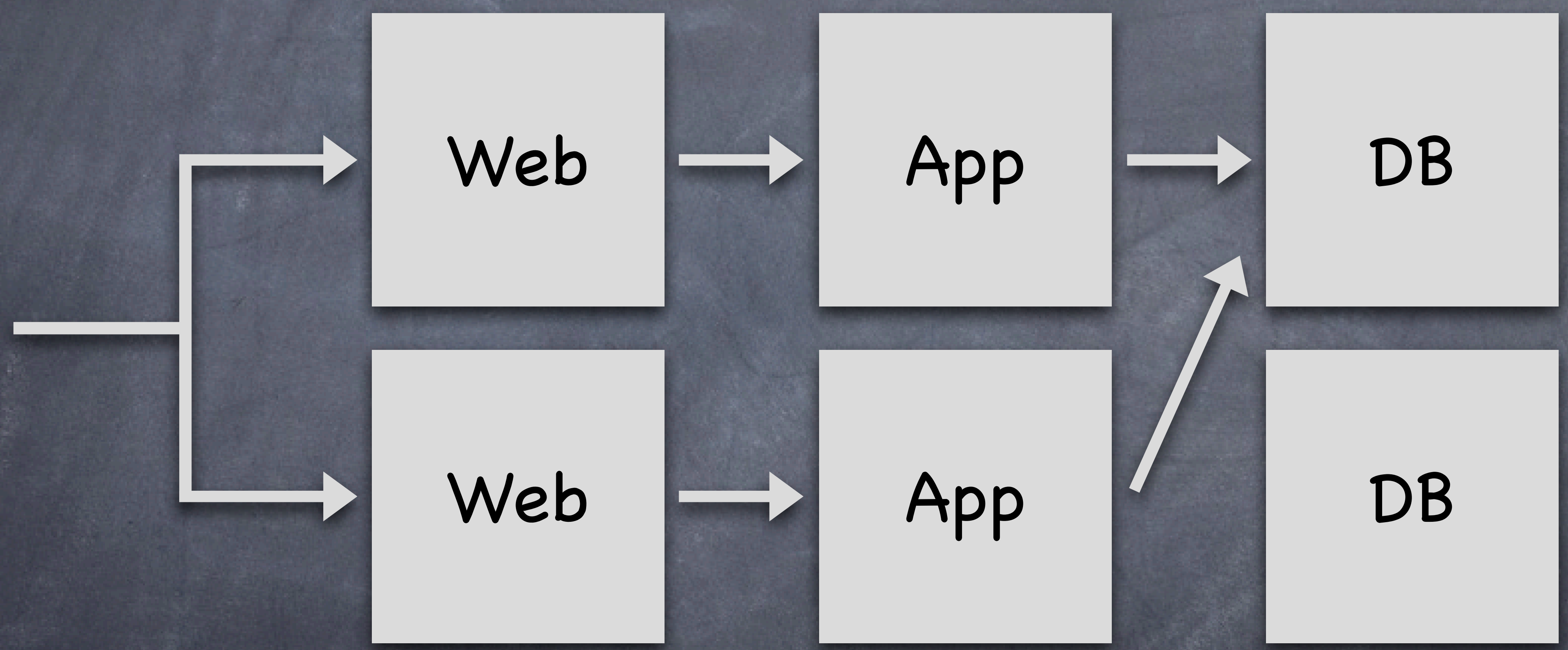
Further assume:
- Identical
- Independent

Reliability with 7 of 9




Parallel
Series
Reliability
Time
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9

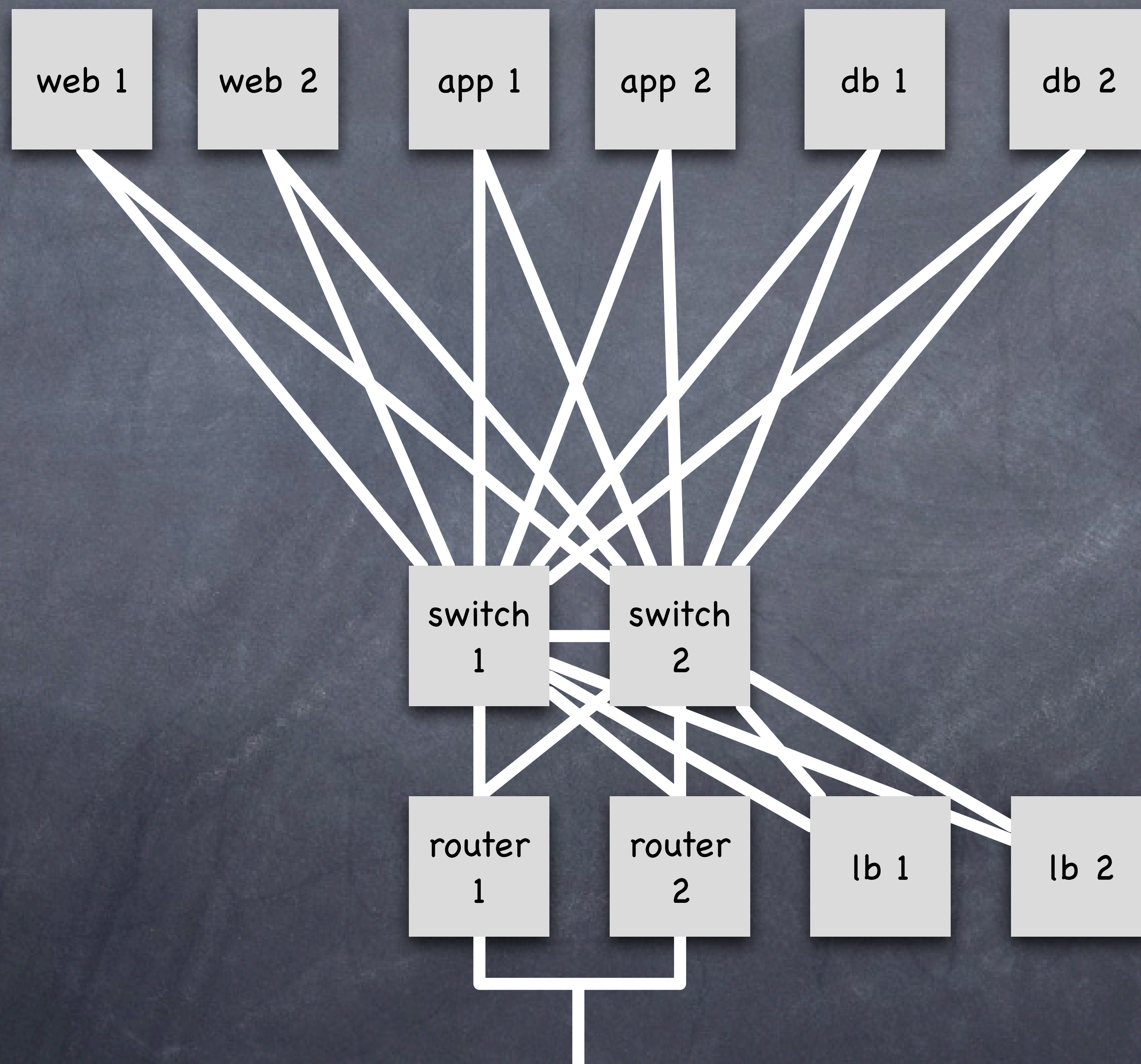
Getting More Real



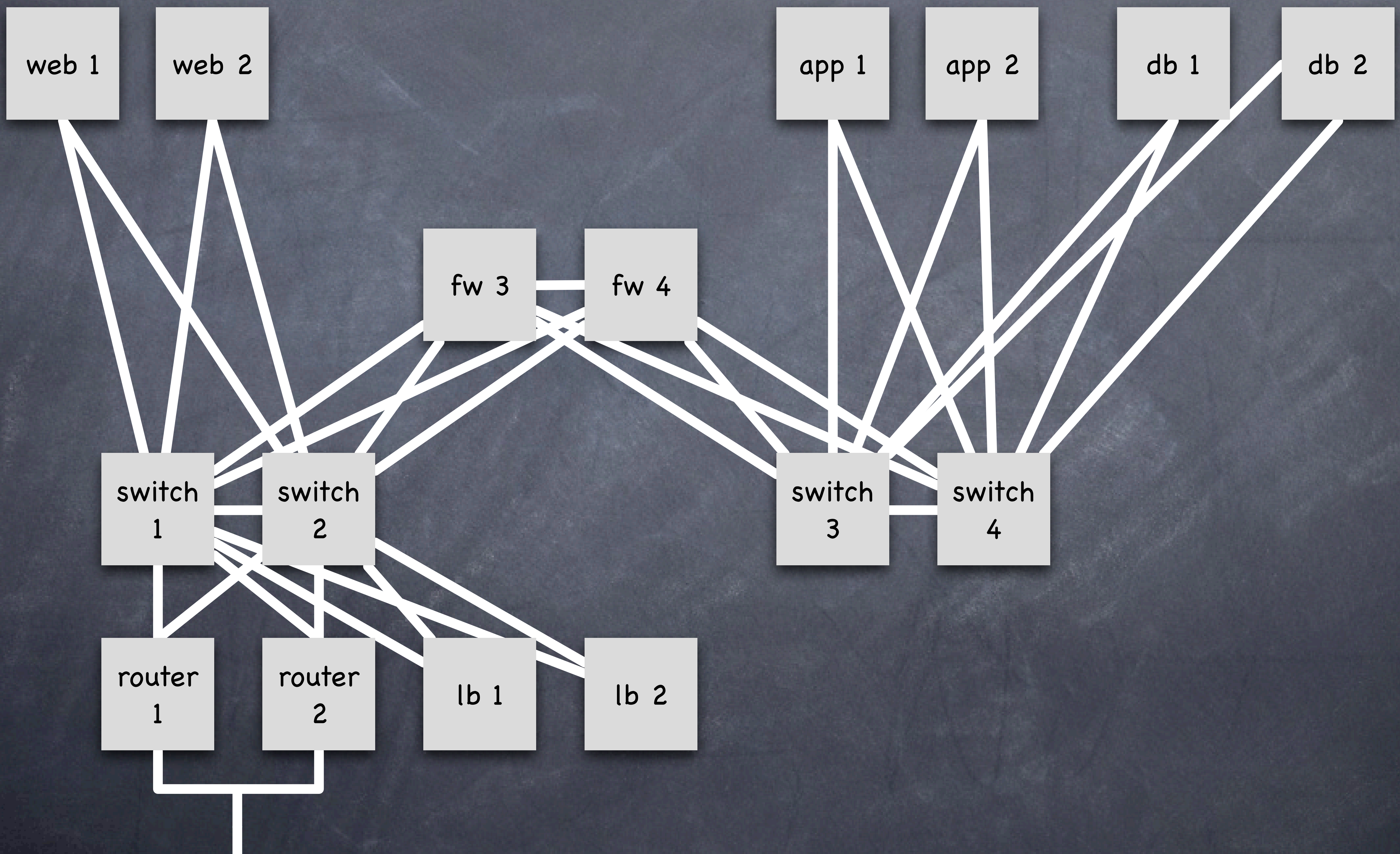
Assume parallel reliability at each layer.

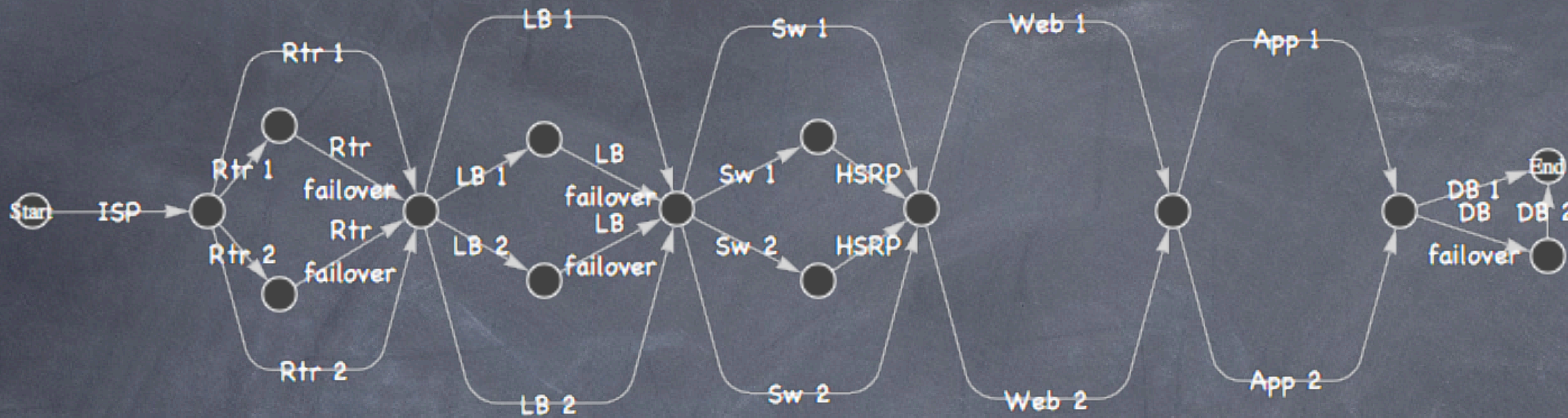
What's missing?

With the Network



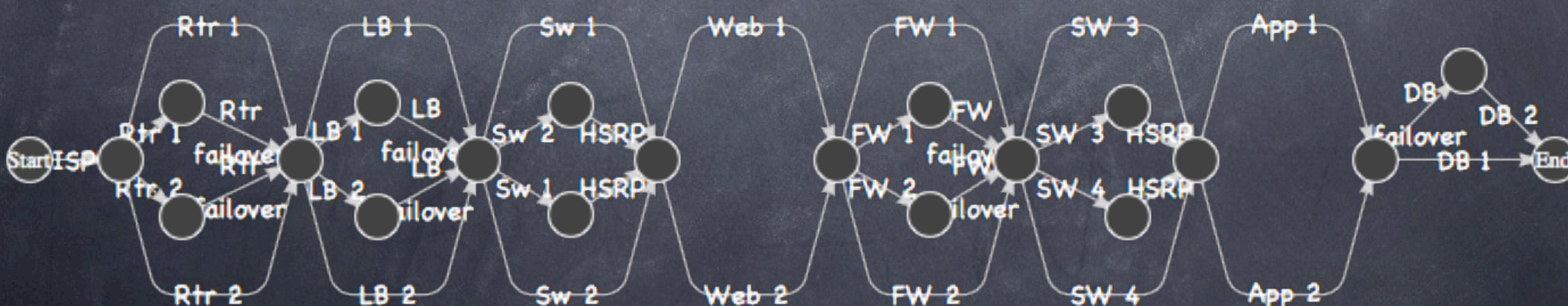
Alternate Network Design



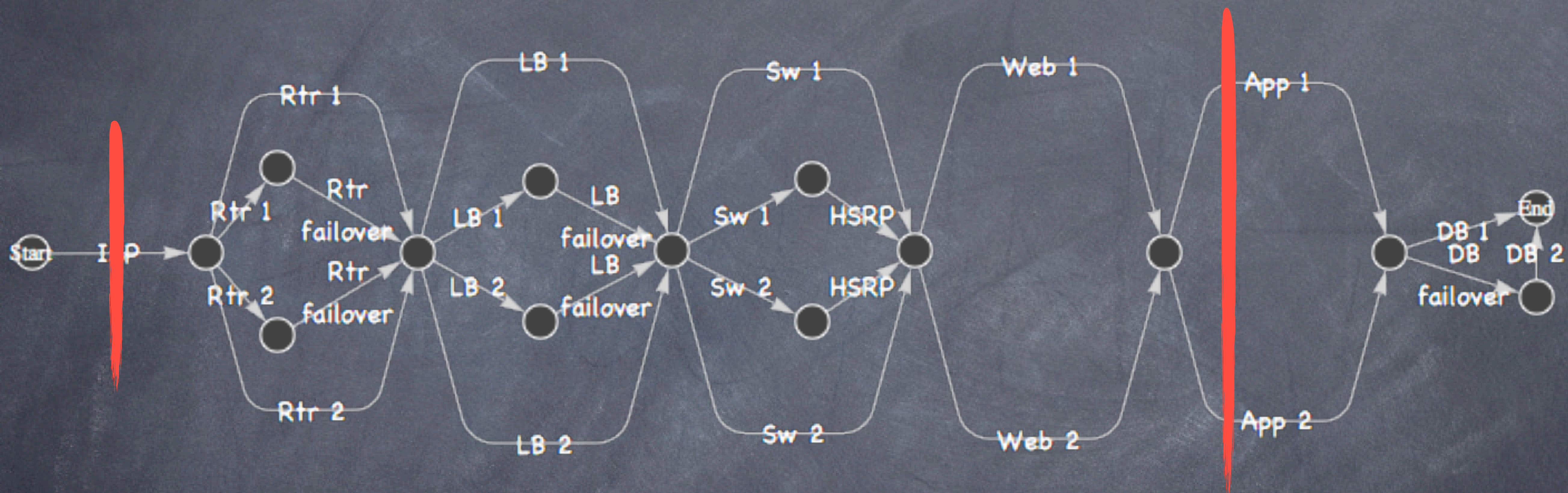


Single Network

Split Network



Using a Reliability Graph



Cut Sets

Cut set of size 1 = S.P.O.F.

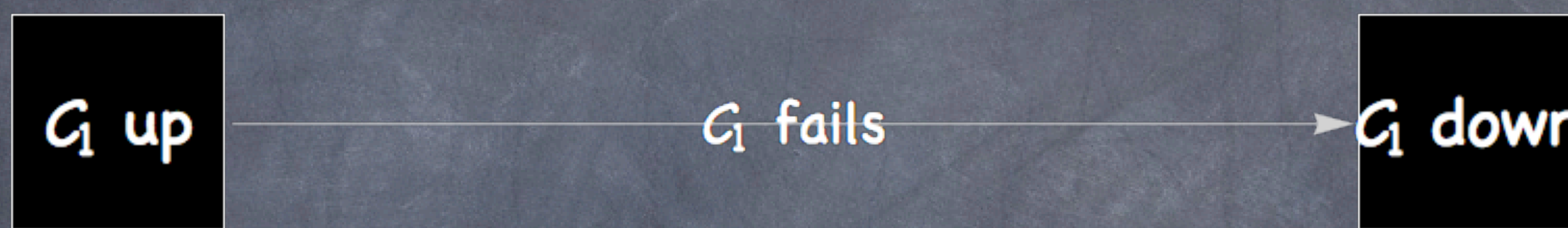
"If a builder build a house for a man and do not make its construction firm and the house which he has built collapse and cause the death of the owner of the house, that builder shall be built to death."

Code of Hammurabi

or

Corporate America

One Component without Repair

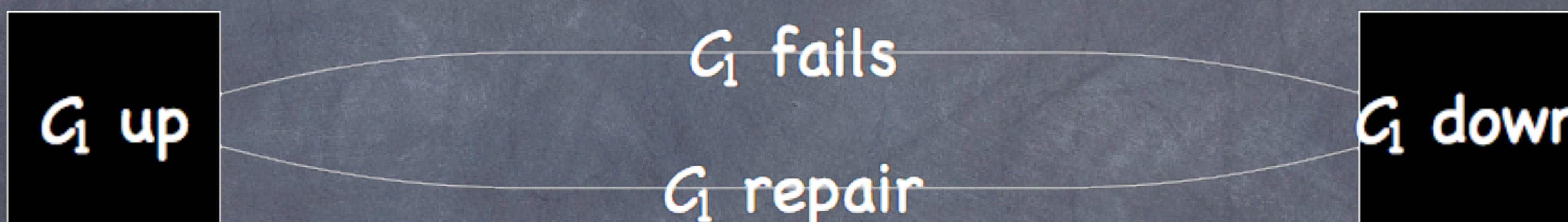


$$P_{failure}(t) = 1 - R_c(t)$$

Assumes that failure is fatal and permanent.

One Component with Repair

$$P_{failure}(t) = 1 - R_c(t)$$



$$P_{repair}(t)$$

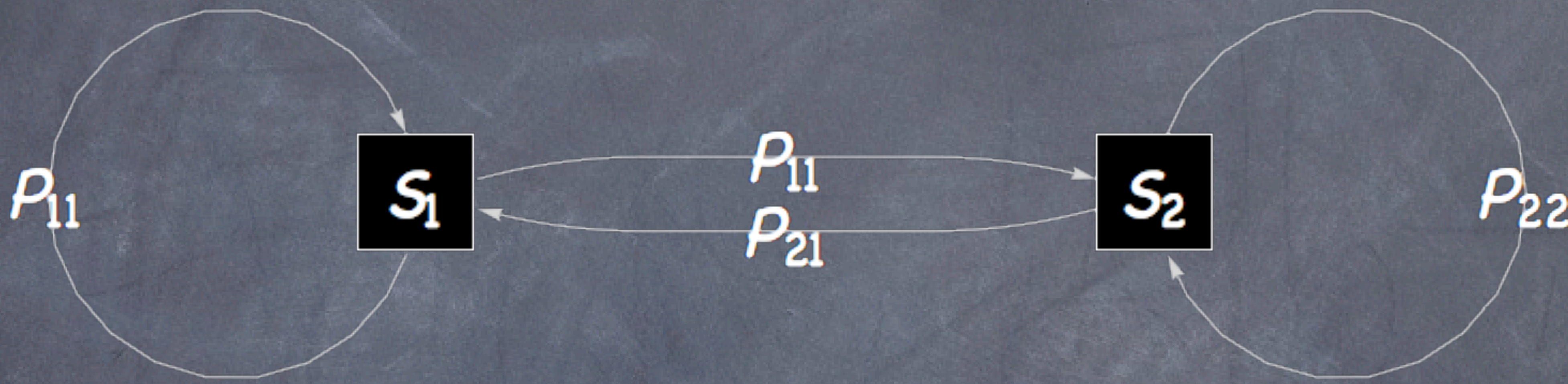
Markov

Not this guy.

Andrey
Markov

This guy.

Markov Process



State i active at any time

Transition from i to j with probability P_{ij}

The system has no memory.

“Easy” Solutions to Markov Processes

$$\frac{dP_{s_0}}{dt} = -[z_{01}(t) + z_{02}(t)]\,P_{s_0}(t)$$

$$R(t) = e^{-(\lambda_1+\lambda_2)t} + \frac{\lambda_1}{\lambda_1+\lambda_2-\lambda_3}\left(e^{-\lambda_3 t} - e^{-(\lambda_1+\lambda_2)t}\right) + \frac{\lambda_2}{\lambda_1+\lambda_2-\lambda_4}\left(e^{-\lambda\right.$$

$$\frac{dP_{s_1}}{dt} = -z_{13}(t)P_{s_1}(t) + z_{01}(t)P_{s_0}(t)$$

$$\frac{dP_{s_2}}{dt} = -z_{23}(t)P_{s_2}(t) + z_{02}(t)P_{s_0}(t)$$

$$\frac{dP_{s_3}}{dt} = z_{13}(t)P_{s_1}(t) + z_{23}(t)P_{s_2}(t)$$

$$z_{01}(t) = \lambda_1, z_{02}(t) = \lambda_2, z_{13}(t) = \lambda_3, z_{23}(t) = \lambda_4$$

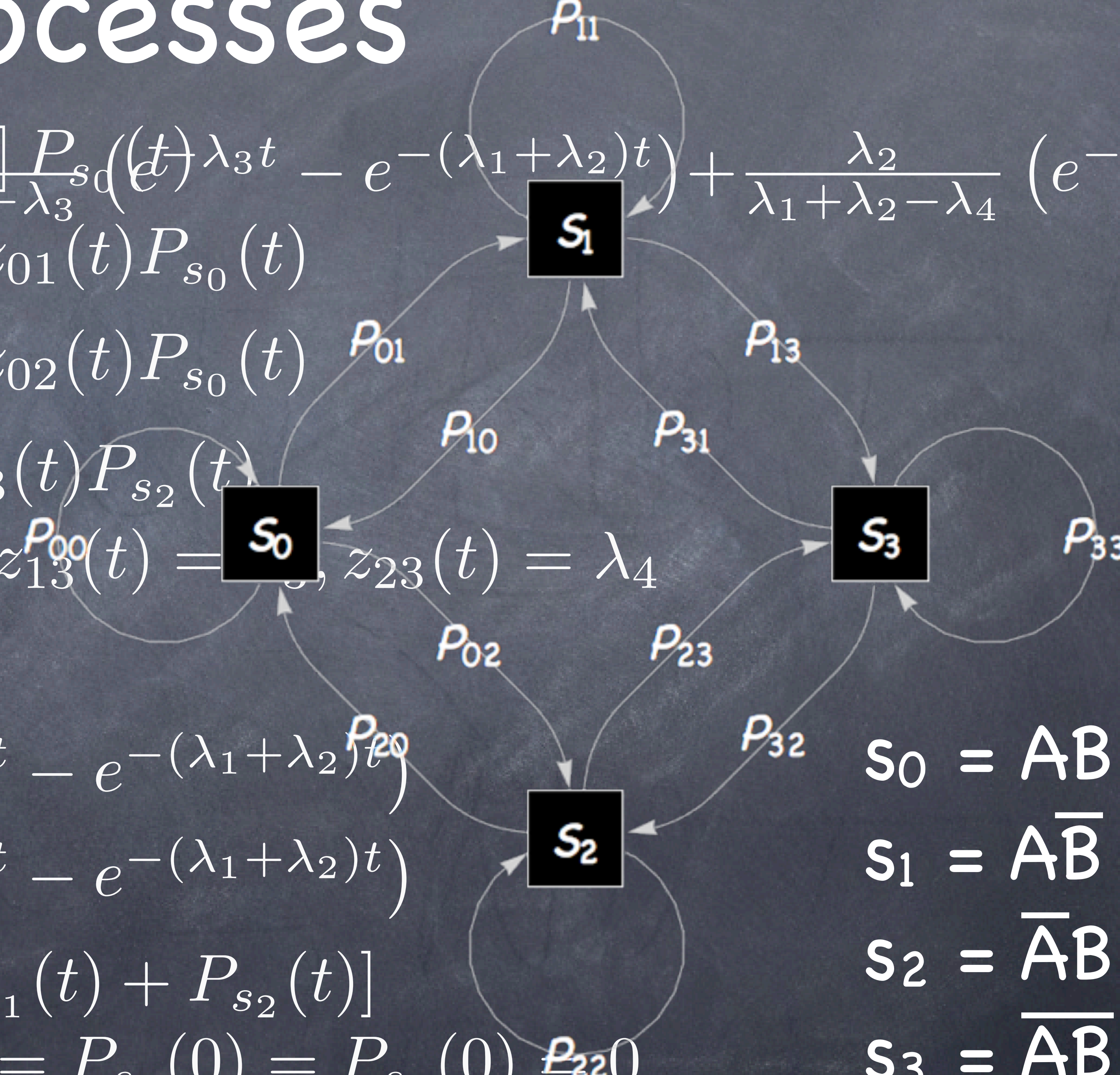
$$P_{s_0}(t) = e^{-(\lambda_1+\lambda_2)t}$$

$$P_{s_1}(t) = \frac{\lambda_1}{\lambda_1+\lambda_2-\lambda_3}\left(e^{-\lambda_3 t} - e^{-(\lambda_1+\lambda_2)t}\right)$$

$$P_{s_2}(t) = \frac{\lambda_2}{\lambda_1+\lambda_2-\lambda_4}\left(e^{-\lambda_4 t} - e^{-(\lambda_1+\lambda_2)t}\right)$$

$$P_{s_3}(t) = 1 - [P_{s_0}(t) + P_{s_1}(t) + P_{s_2}(t)]$$

$$P_{s_0}(0) = 1; P_{s_1}(0) = P_{s_2}(0) = P_{s_3}(0) = 0$$



$s_0 = AB$

$s_1 = A\overline{B}$

$s_2 = \overline{A}B$

$s_3 = \overline{AB}$

1. Find the hazard functions for transitions
2. Write the differential equations
3. Apply the Laplace transform to all equations
4. Solve algebraically in transform domain
5. Apply inverse Laplace transform

Sounds better,
sort of.

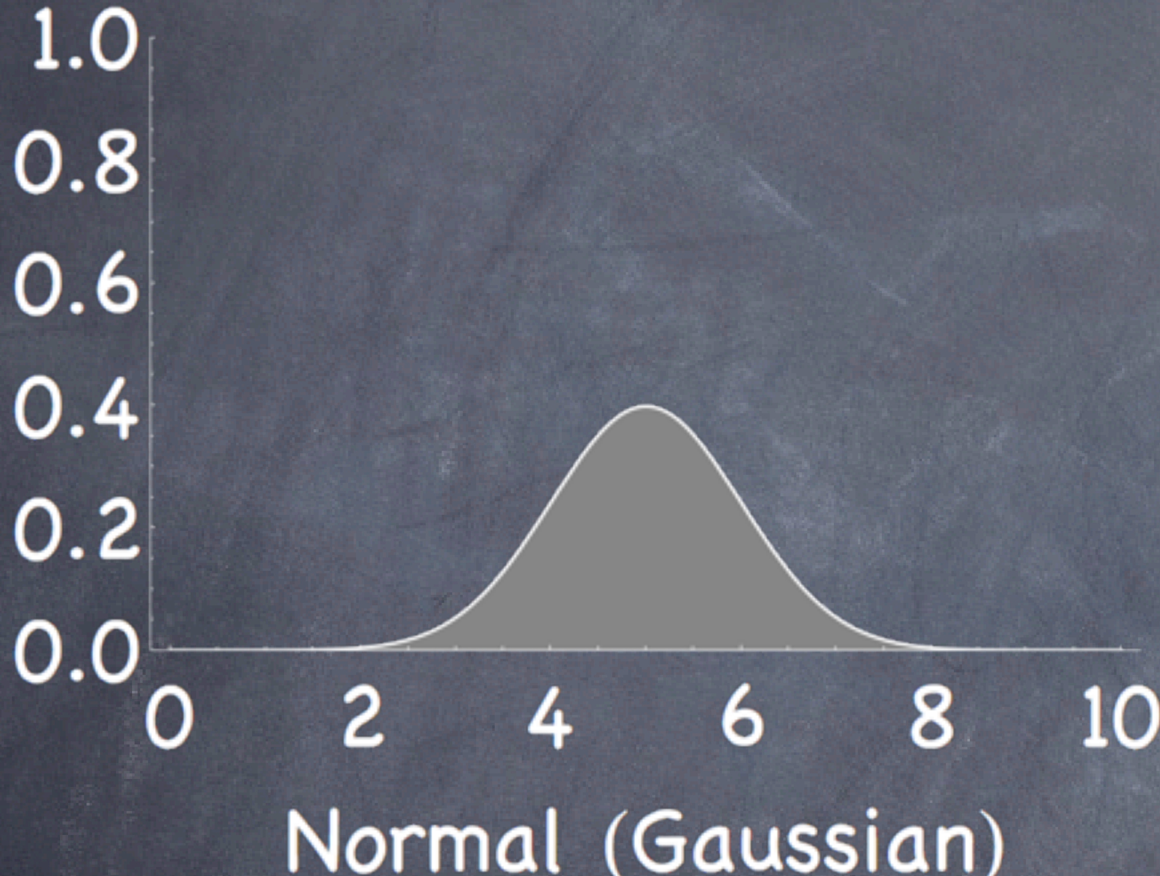
Analytic solutions have more fundamental problems.

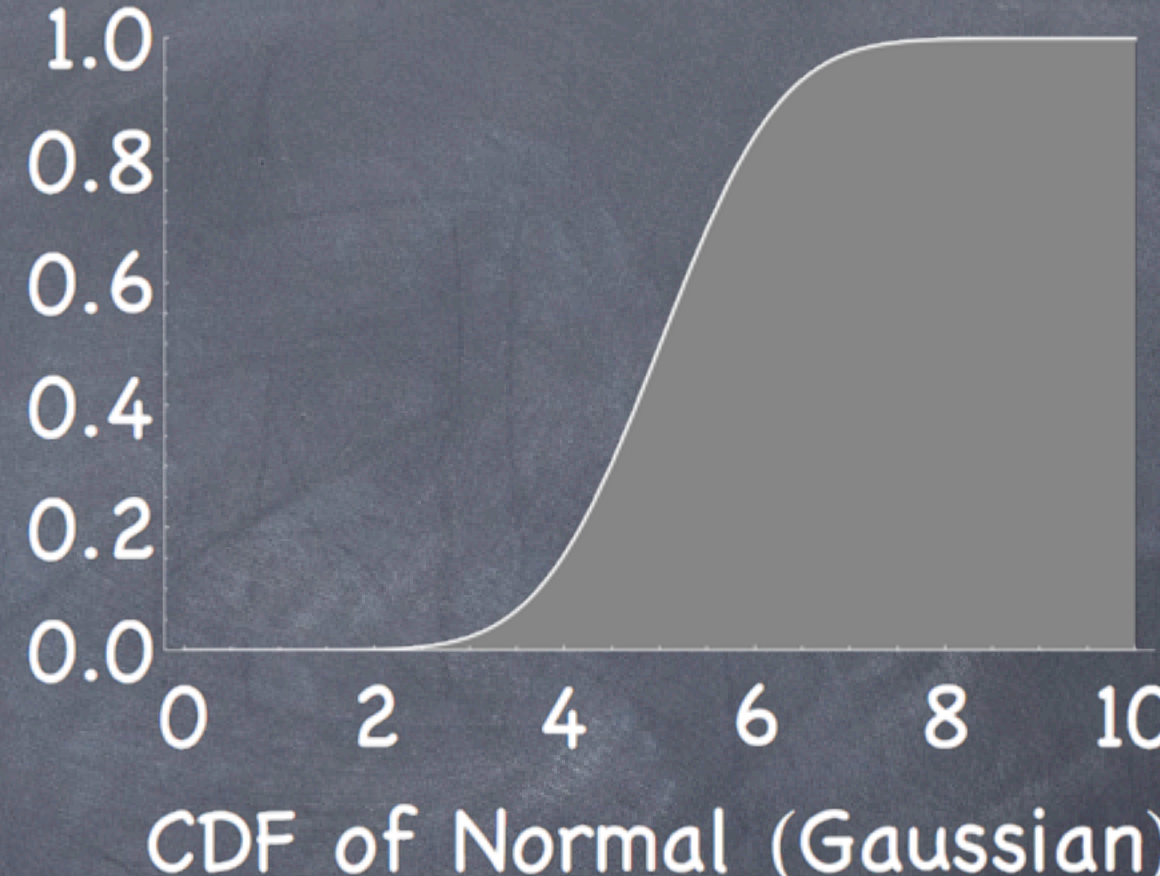
Limitations of Reliability Engineering

Intractable Math

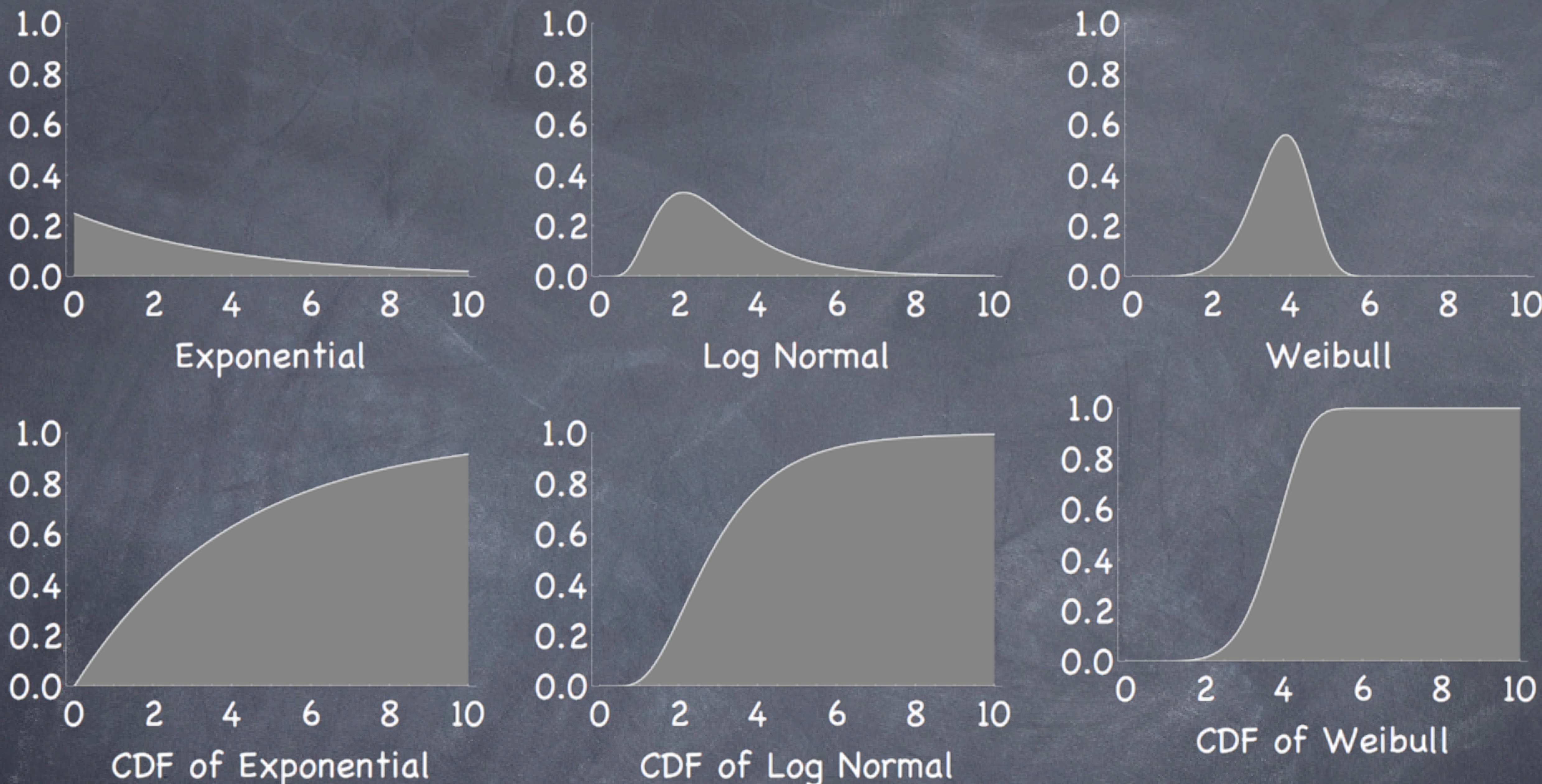
Only some distributions have closed form solutions

Distributions





Density Function $\xrightarrow{\int}$ Cumulative Distribution



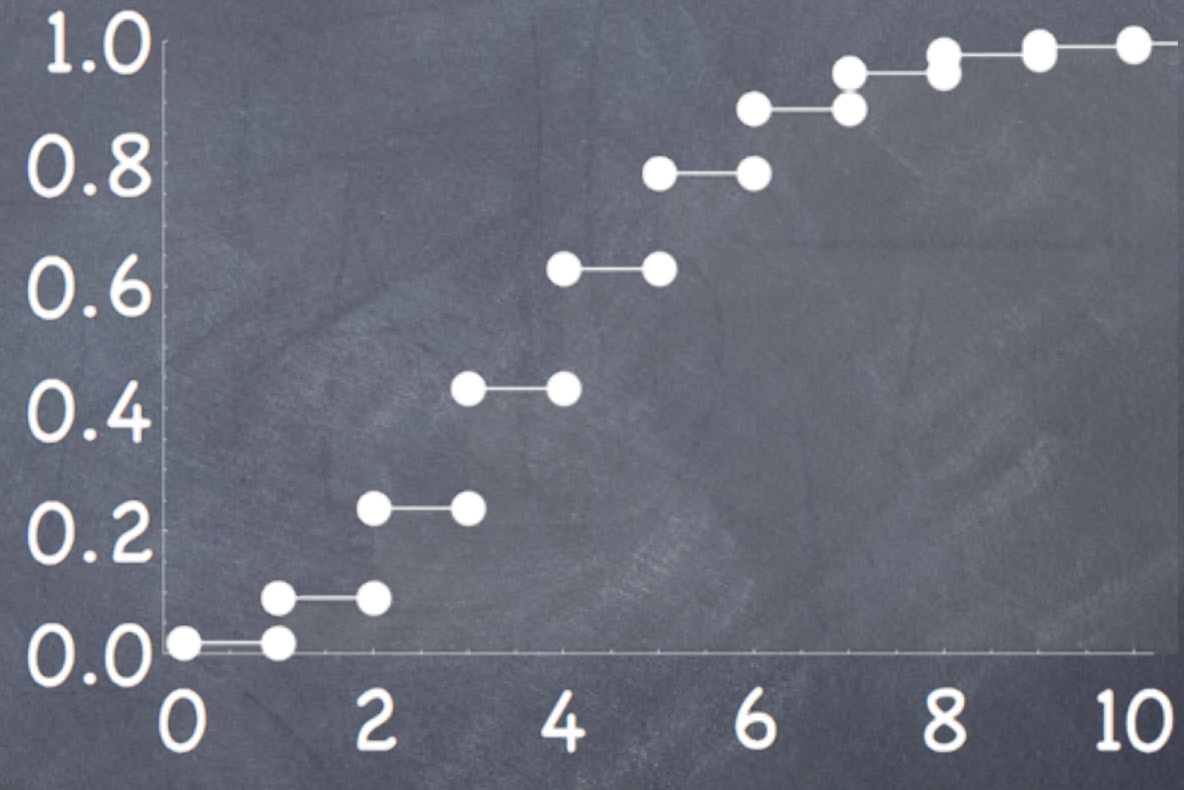
Single component

"Multiplicative" failures

Hardware

Repair Distributions

Poisson



CDF of Poisson

Which of these distributions should we apply to software?

None.

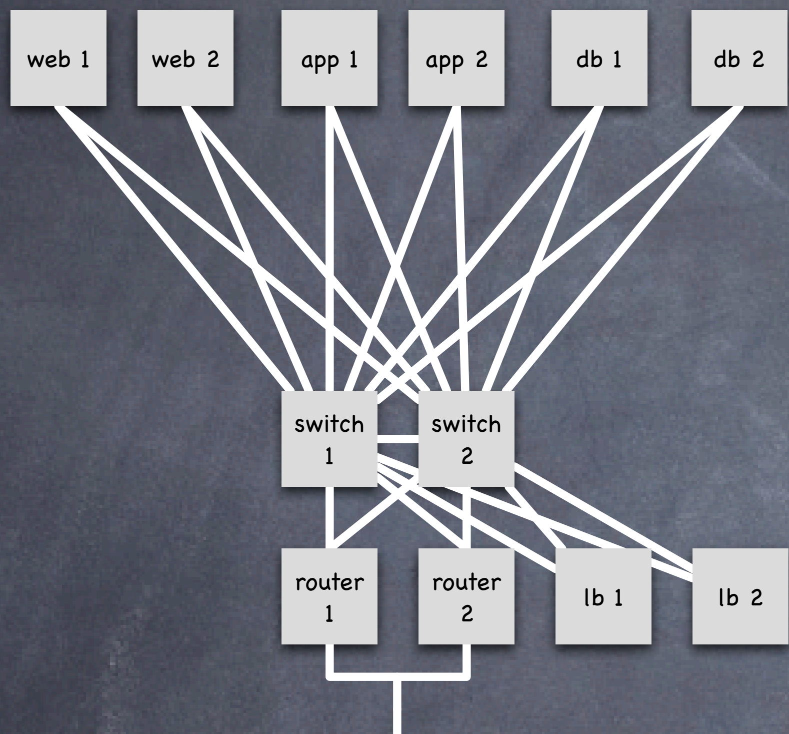
Software fails under on load,
not over time.

Curse of Dimensionality

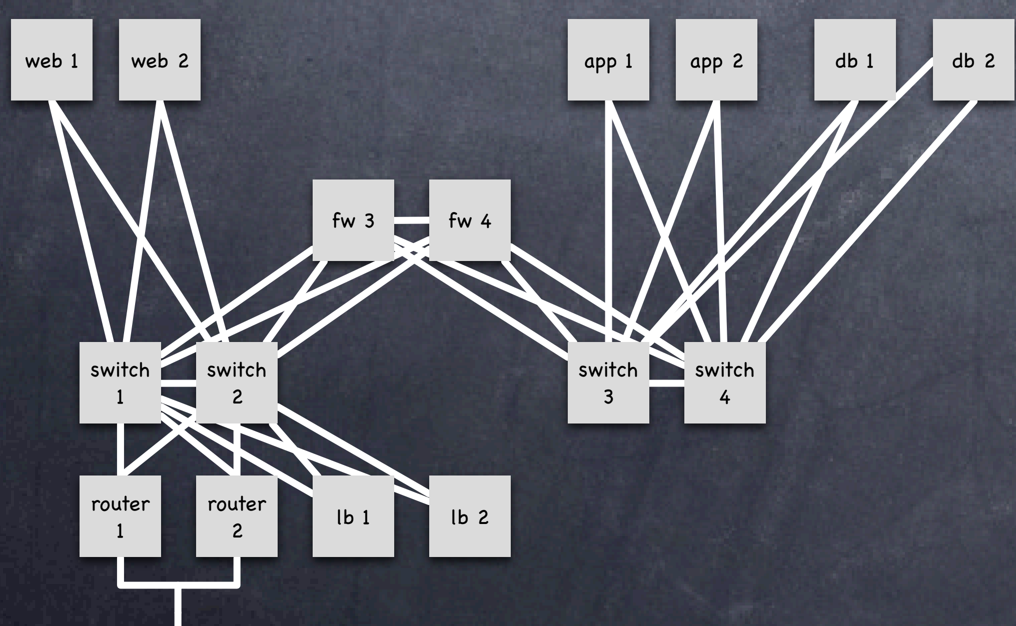
Each fallible component adds a dimension to the Markov model.

States = $O(2^N)$

Size of Transition Matrix



131,072 x 131,072



8,388,608 x 8,388,608

Other Challenges

Failure distributions are a deep, dark secret

MTBF == BS

How do you measure MTBF?

1. Get a disk drive
2. Run random reads & writes
3. See how long it keeps running

Not really.

Livermore-Pleasanton Fire Station #6

110 years

Testing one device doesn't tell you much about the population.

Ergodic Hypothesis

Behavior of a population approximates behavior over time.

Ergodic MTBF

Test 10 units for 500 hours, under an acceleration model. Count failures, model failure rate under real use conditions.

Other big killers

- Human error: 50 - 65% of outages
- Interiority
- Distributed failure modes
- Lack of independence between nodes & layers

Should we abandon reliability engineering?

NO!

Why go to the trouble?

R.E. cannot say you're OK.

But it can say you're definitely **not** OK.

Cost vs. Benefit

Modeling reduces uncertainty.

Use RE like other models.

Apply when your system is at risk in that dimension.

Reliability Engineering Matters

(Except when it doesn't)

Michael T. Nygard
Relevance, Inc.

michael.nygard@thinkrelevance.com @mtnygard