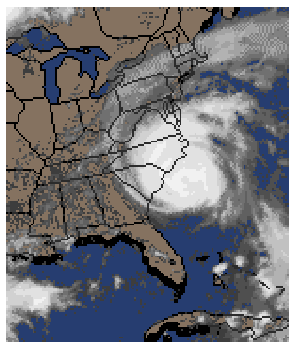
Essential Performance Forecasting

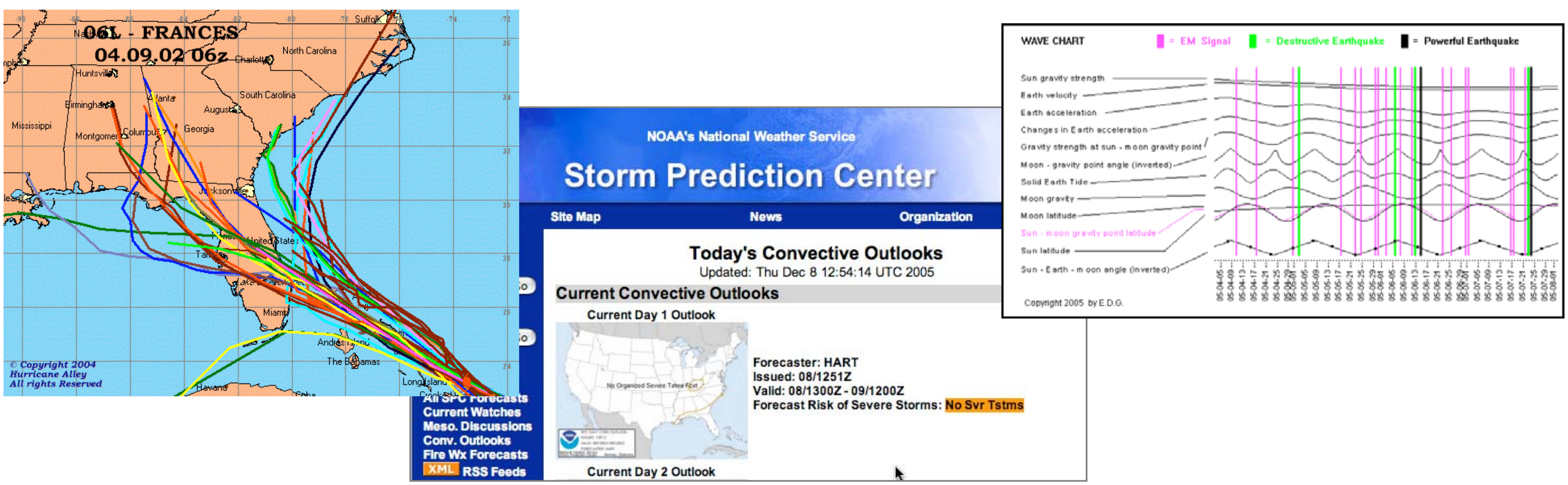


NYOUG - December 2005

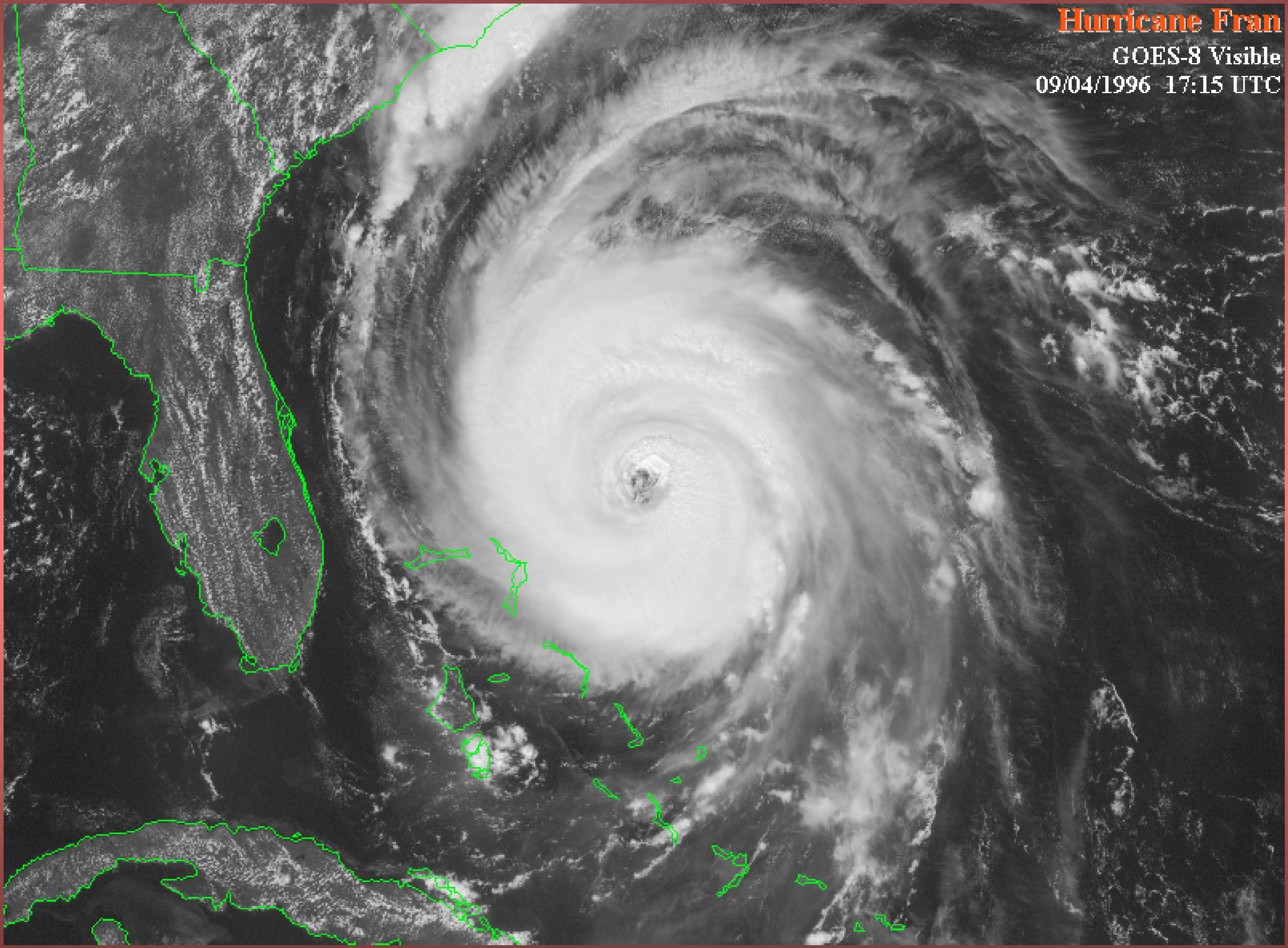
Craig A. Shallahamer - craig@orapub.com

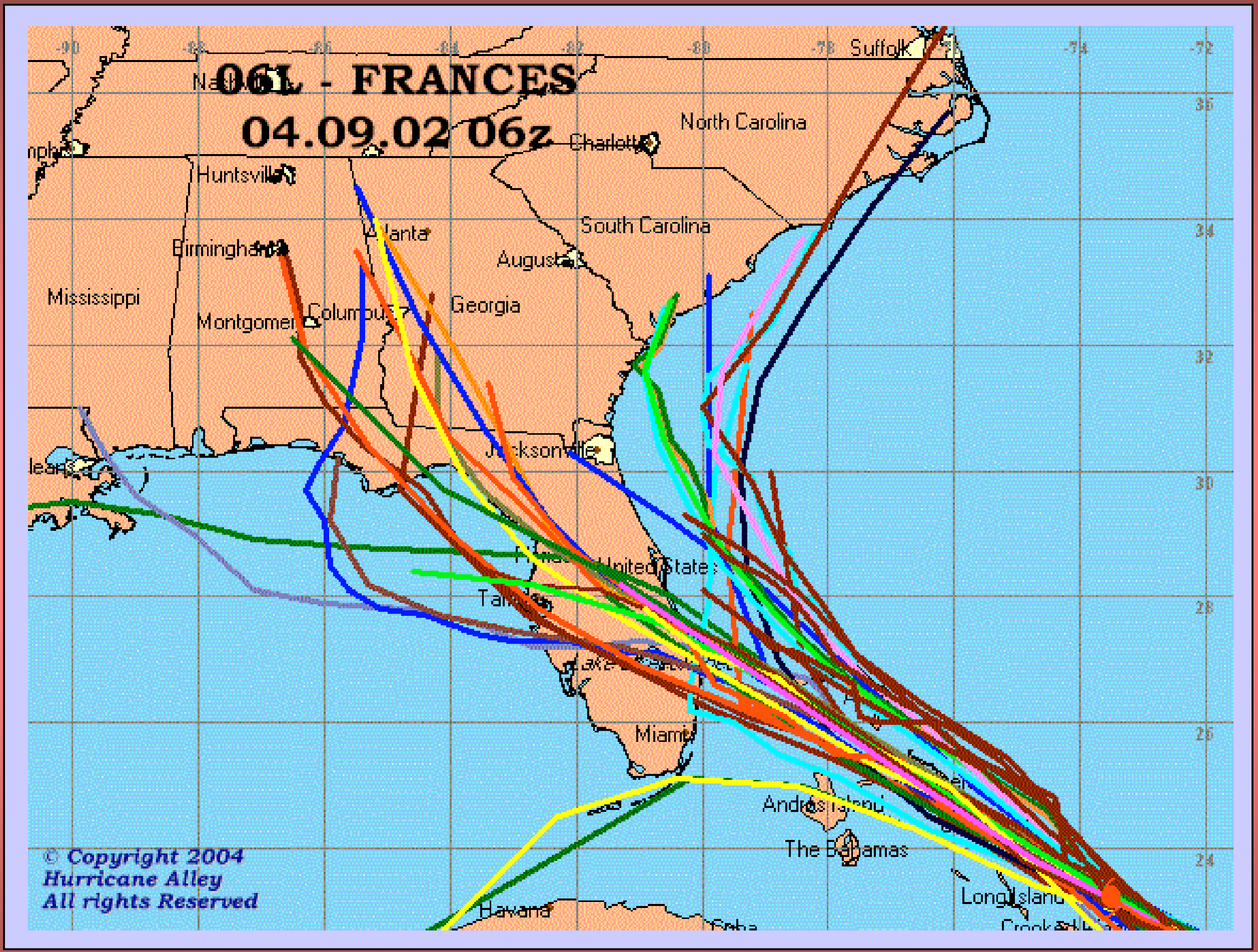
Forecasting is not about numbers…

- It's about predictions…
- It's about foretelling the future…
- And the world needs people who can do this stuff!



We got hurricanes.





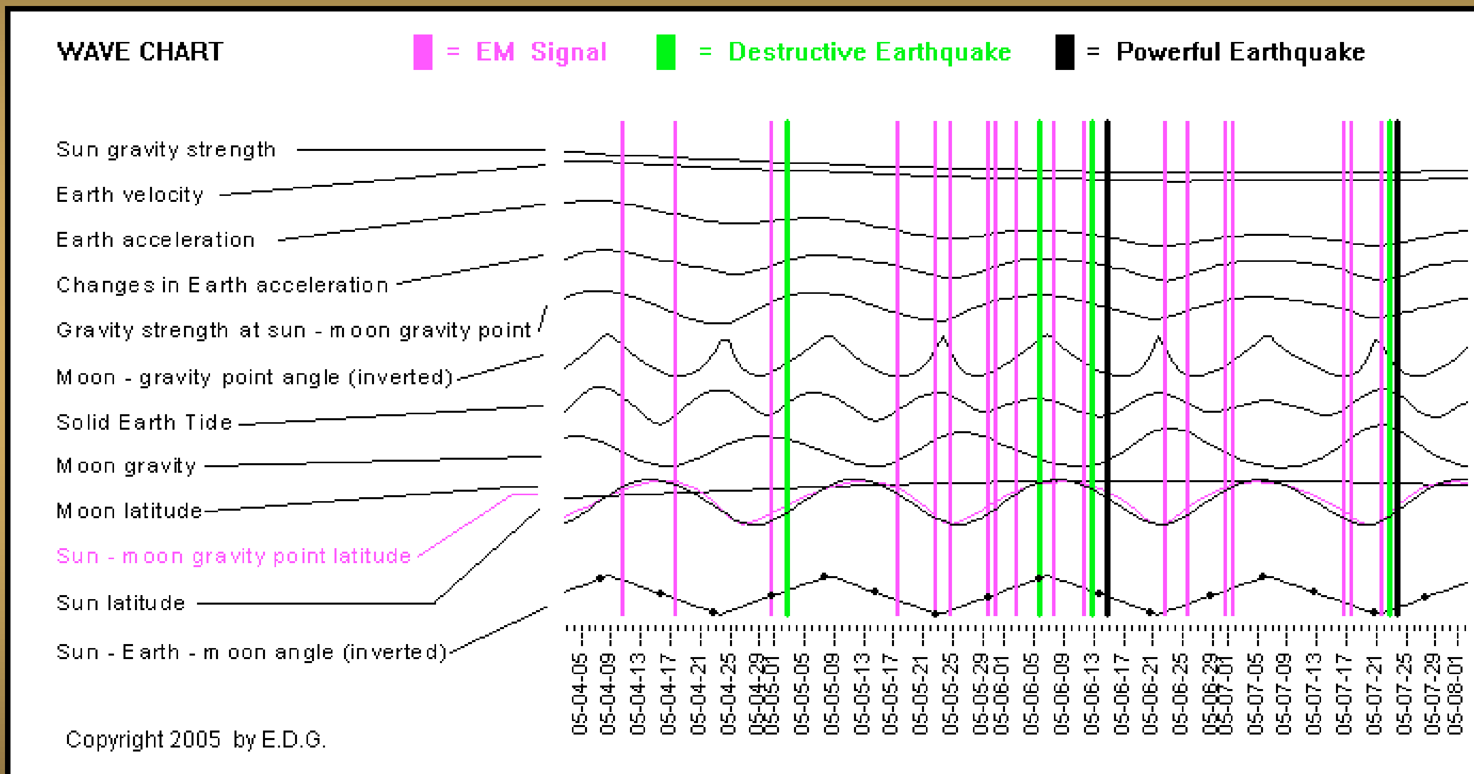
We need forecasting.

We got tornadoes.

We need forecasting.

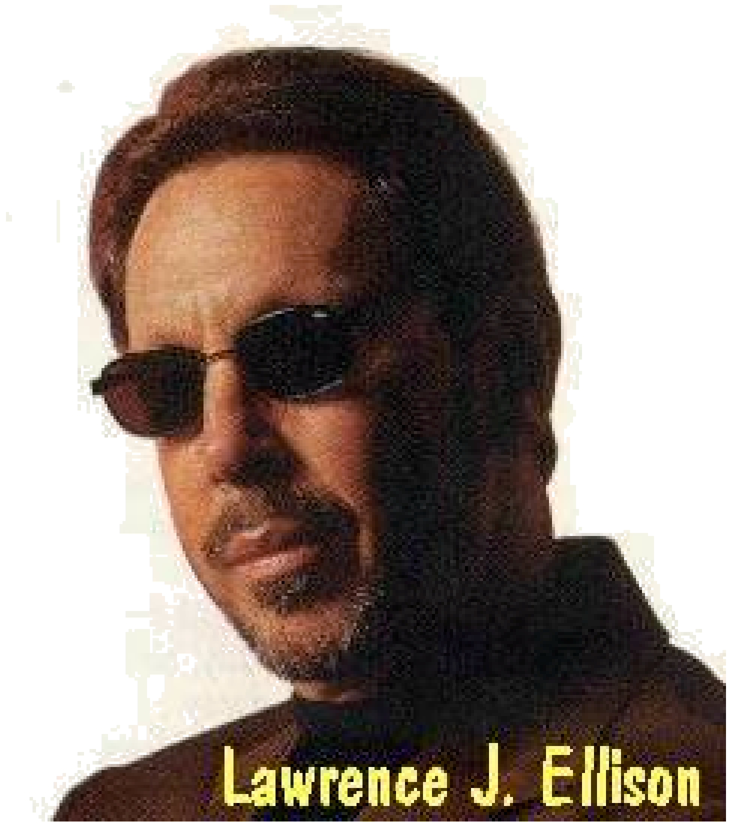
We got earthquakes.

We need forecasting.

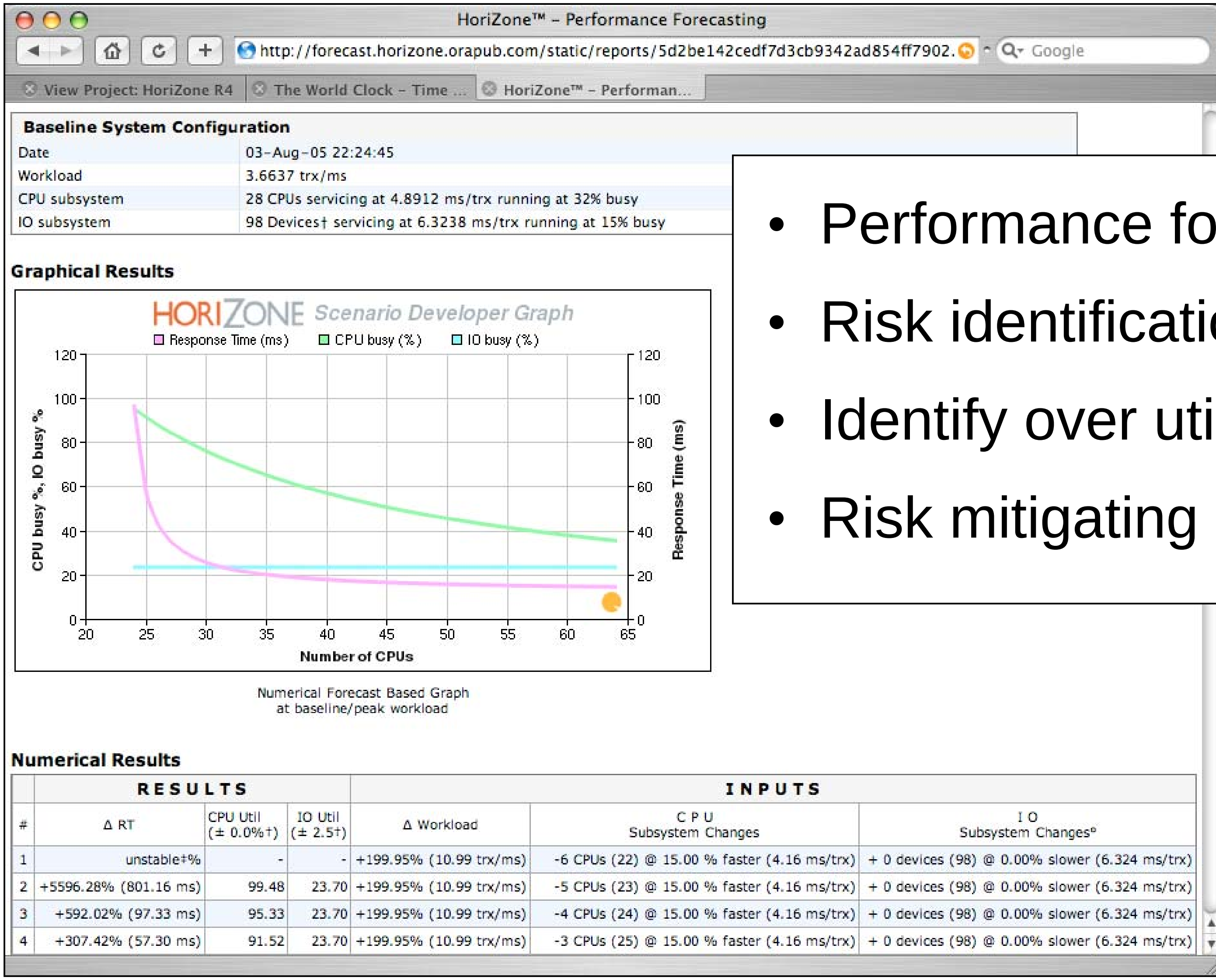


We got a 64 CPU HP Superdome with Oracle 10g Enterprise Edition retailing at $2,500,000.

We need…



We need forecasting…



#	Δ RT	CPU Util (± 0.0%†)	IO Util (± 2.5†)	Δ Workload	CPU Subsystem Changes	IO Subsystem Changes°
1	unstable‡%	-	-	+199.95% (10.99 trx/ms)	-6 CPUs (22) @ 15.00 % faster (4.16 ms/trx)	+ 0 devices (98) @ 0.00% slower (6.324 ms/trx)
2	+5596.28% (801.16 ms)	99.48	23.70	+199.95% (10.99 trx/ms)	-5 CPUs (23) @ 15.00 % faster (4.16 ms/trx)	+ 0 devices (98) @ 0.00% slower (6.324 ms/trx)
3	+592.02% (97.33 ms)	95.33	23.70	+199.95% (10.99 trx/ms)	-4 CPUs (24) @ 15.00 % faster (4.16 ms/trx)	+ 0 devices (98) @ 0.00% slower (6.324 ms/trx)
4	+307.42% (57.30 ms)	91.52	23.70	+199.95% (10.99 trx/ms)	-3 CPUs (25) @ 15.00 % faster (4.16 ms/trx)	+ 0 devices (98) @ 0.00% slower (6.324 ms/trx)

- Performance forecasting…
- Risk identification…
- Identify over utilized resources…
- Risk mitigating strategies…

So when your boss comes up to you on a Friday afternoon and asks,

"On Monday the new subsidiary is going to be added to our system. That's not going to be a problem, is it?"

Do you sit paralyzed in fear…
Knowing that the career opportunity of your life just passed before you…
?

Or… Do you…

- Forecast response time change…
- Identify risk…
- Forecast over utilized resources…
- Develop risk mitigating strategies…

So what's your next move?

Here's how to get started.

- Basic understanding of computing systems.
- Basic queuing theory understanding.
- Basic math.
- Workload data.
- Understanding how to put all this stuff together.

…the essentials of forecasting Oracle performance.

Our love for predicting the future…

Even before recorded history…

A good prediction identifies risk and provides insights

A computing system is alive!

- Computing systems are like living systems. Think: honeybee colony or the Earth's water cycle.

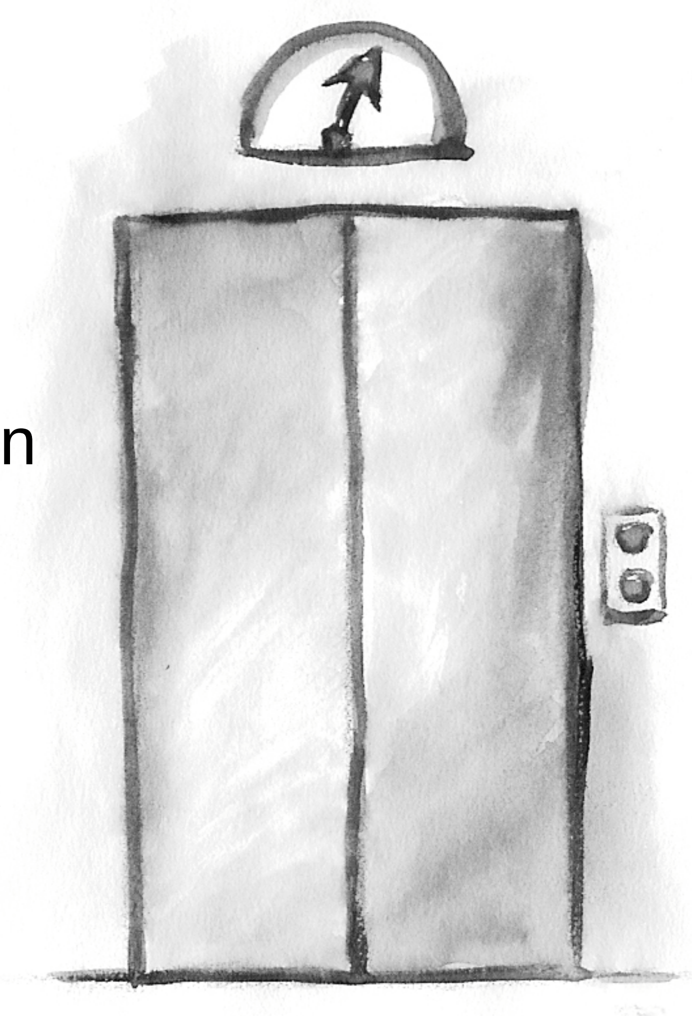
- Systems need:
  - Energy
  - Guidelines

The arrival rate : λ

- Transactions arrive into a computing system like people arrive into an office building.

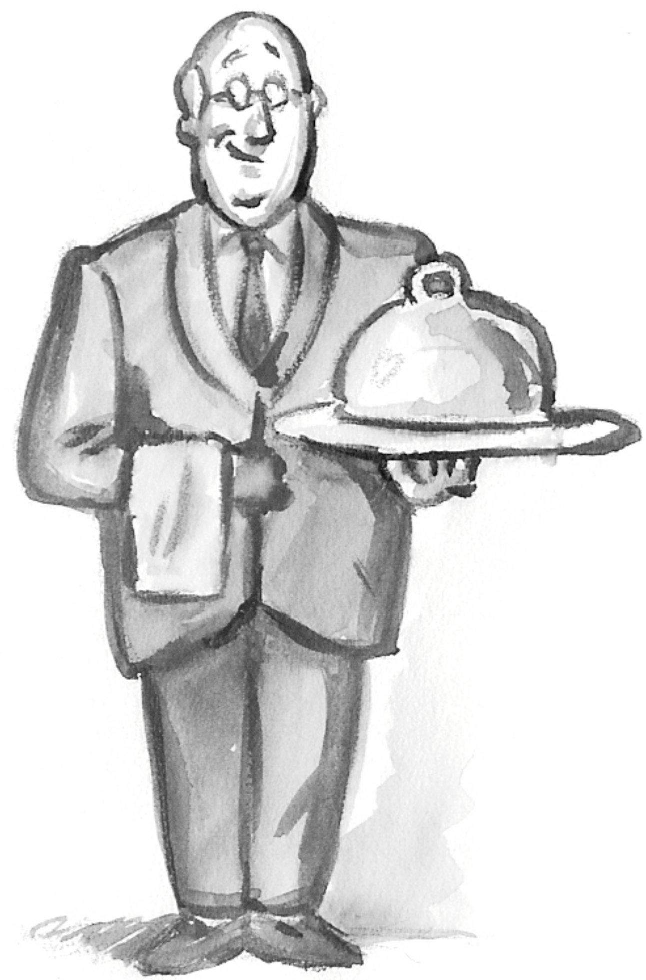
- There are many statistics we can use to measure the arrival rate.

- Common statistics from **v$sysstat**; logical reads, blocks changes, physical writes, user calls, logons, executes, user commit, and user rollbacks.



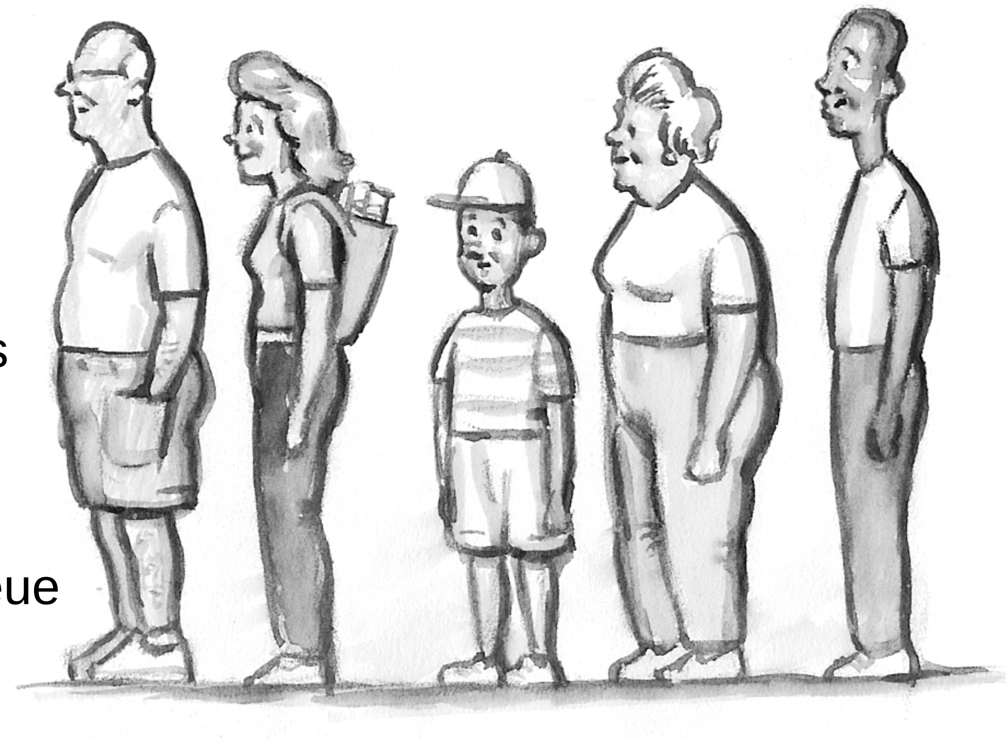
The transaction processor…server.

- "How can we serve you?" That person is a server.
- CPU and IO devices are servers.
- Each transaction consumes service time, **S**.
- The service time is how long it takes a server to process a transaction.
- The busyness of a server is called the utilization, **U**.
- When a server gets above 70% utilized, transactions start to wait.



The queue.

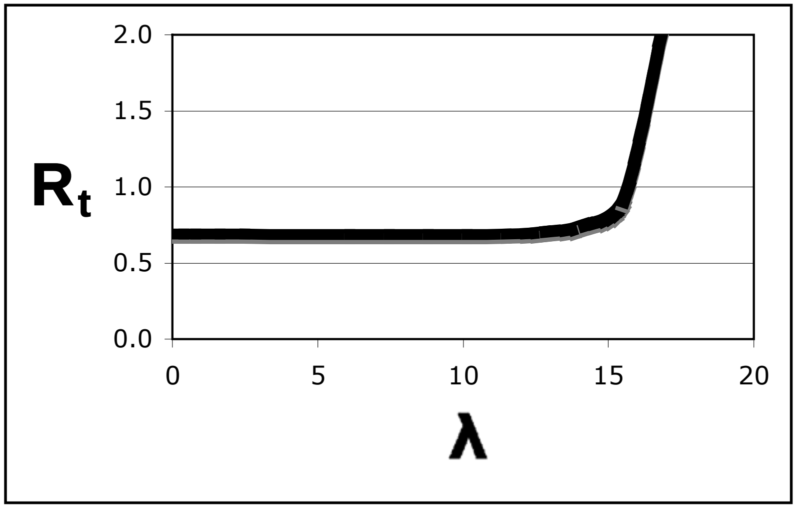
- Ever been told to wait while a hostess writes down your name? You were placed into the queue!
- When a transaction waits, it is placed into a queue.
- Each queue has a length, **Q**.
- Each transaction is in the queue for time **W**.

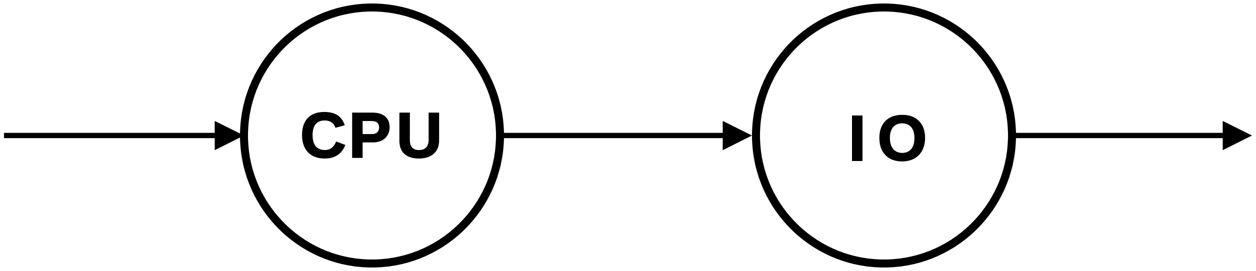


- Performance decreases when a server gets busy and transactions queue.
- This occurs at around 75%.

Transactions flow.

- When a business transaction is submitted, it flows throughout the computing system,
- Consuming CPU, IO, memory, and network resources.
- A transaction may have to queue before securing a server.
- The sum of service time and queue time is called response time, **R**.
- The more detailed the forecast model, the more we detail transaction flow.
- This does not mean the forecast is more precise.





Steps to forecast performance.

- **Determine the study question**. What is the question we must answer.
- **Characterize the workload**. Gather data and appropriately format.
- **Develop and use appropriate model**. Pick the "best" model.
- **Validate forecast**. Ensure the forecast is working and understand its precision. Decide if it's appropriate to forecast with.
- **Forecast**. Actually do the forecasting.

Gathering performance data.

- Before we can forecast, we must gather performance data.
- We can gather from a proposed, benchmarked, or production system.
- Make sure data from different subsystems is gathered at the same time.

```
$ sar -u 300 1
SunOS soul 5.8 Generic_108528-03
10:51:00  %usr  %sys  %wio %idle
10:56:00    27     8     0    65
```

```
SQL> select name, value
  2  from v$sysstat
  3  where name='user calls';

NAME              VALUE
------------ ----------
user calls      5006032

SQL> /
NAME              VALUE
------------ ----------
user calls      5007865
```

Doing the math.

- λ : Arrival rate (trx/sec)
- S : Service time (sec/trx)
- U : Utilization
- Q : Queue length
- W : Wait time
- R : Response time
- M : Number of servers

Basic CPU Formulas

```
U = ( S λ ) / M
R = S / ( 1 - U^M )
Q = ( MU / ( 1 - U^M ) ) - M
```

Basic IO Formulas

```
U = ( S λ ) / M
R = S / ( 1 - U )
R = S + W
```

Understanding the math.

- What happens to CPU utilization when the arrival rate increases?
- What happens to CPU utilization when we use faster CPUs?
- What happens to CPU response time when we use faster CPUs?
- What happens to CPU response time if utilization increases?
- What happens to IO response time if service time decreases?
- What happens to IO response time if we increase the number of devices?

Basic CPU Formulas

$U = (S \lambda) / M$

$R = S / (1 - U^M)$

$Q = (MU / (1 - U^M)) - M$

Basic IO Formulas

$U = (S \lambda) / M$

$R = S / (1 - U)$

$R = S + W$

A real life CPU example.

- From our data gathered:
  - 12 CPUs
  - CPU utilization is 35%.
  - Arrival rate is 6.11 uc/sec.
- So…
  - S = 0.69 sec/uc
  - R = 0.69 sec/uc
  - Q = 0
- Do you think there is a performance problem?

Basic CPU Formulas

```
U = ( S λ ) / M
R = S / ( 1 - U^M )
Q = ( MU / ( 1 - U^M ) ) - M
```

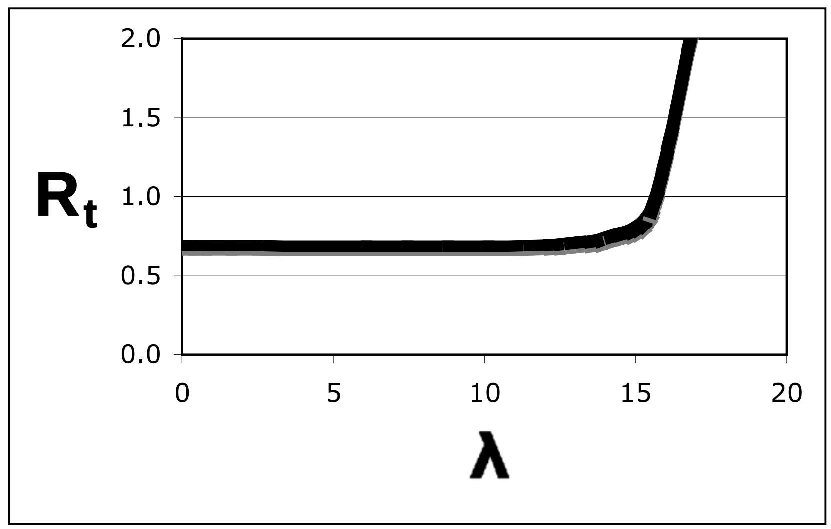
Here's the math.

```
S = UM/λ = 0.69 sec/uc
R = S/(1-U^M) = 0.69 sec/uc
Q = ( MU / (1-U^M) ) - M = -7.8
```

“What if” analysis.

- Powerful forecasting begins when you combine many individual forecasts into a *scenario*.
- Scenario forecasting allows you to create trends and graphs.
- With both numeric and graphical results, you can easily see how a system will respond under different workloads and configurations.
- It’s much easier to identify risk and develop risk mitigating strategies using scenario forecasting.

Inputs		Forecasts		
% Increase	Arrival Rate	Busy	Response Time	Queue Length
0	6.11	0.35	0.69	-7.80
22	7.45	0.43	0.69	-6.88
44	8.80	0.50	0.69	-5.95
66	10.14	0.58	0.69	-5.02
88	11.49	0.66	0.69	-4.05
110	12.83	0.74	0.70	-2.96
132	14.18	0.81	0.75	-1.38
154	15.52	0.89	0.91	2.11
176	16.86	0.97	2.02	22.12



Basic Forecasting Formulas

Basic CPU Formulas

$U = (S \lambda) / M$

$R = S / (1 - U^M)$

$Q = (MU / (1 - U^M)) - M$

Basic IO Formulas

$U = (S \lambda) / M$

$R = S / (1 - U)$

$R = S + W$

U	Utilization
S	Service Time (sec)
λ	Arrival Rate (trx/sec)
M	Number of servers
Q	Queue length
W	Queue time

Case Study : Bob

Bob’s manager Frank (who is actually Bob’s wife’s cousin’s brother’s friend) needs to reduce the cost of their Oracle database license.

Since the database license is based upon the number of CPUs, Frank asked Bob to forecast the change in response time if they were to remove CPUs from their database server.

Bob has repeatedly observed that the 26 CPU HP server is usually around 28% busy during peak processing time (month end close).

Bob's Solution

We are given:

Number of CPUs (M) = 26

Utilization (U) = 28%

Set the arrival rate (λ) to 1.

Now derive the service time (S);

$$U = S \lambda / M$$

$$S = UM\lambda = .28 * 26 * 1 = 7.28$$

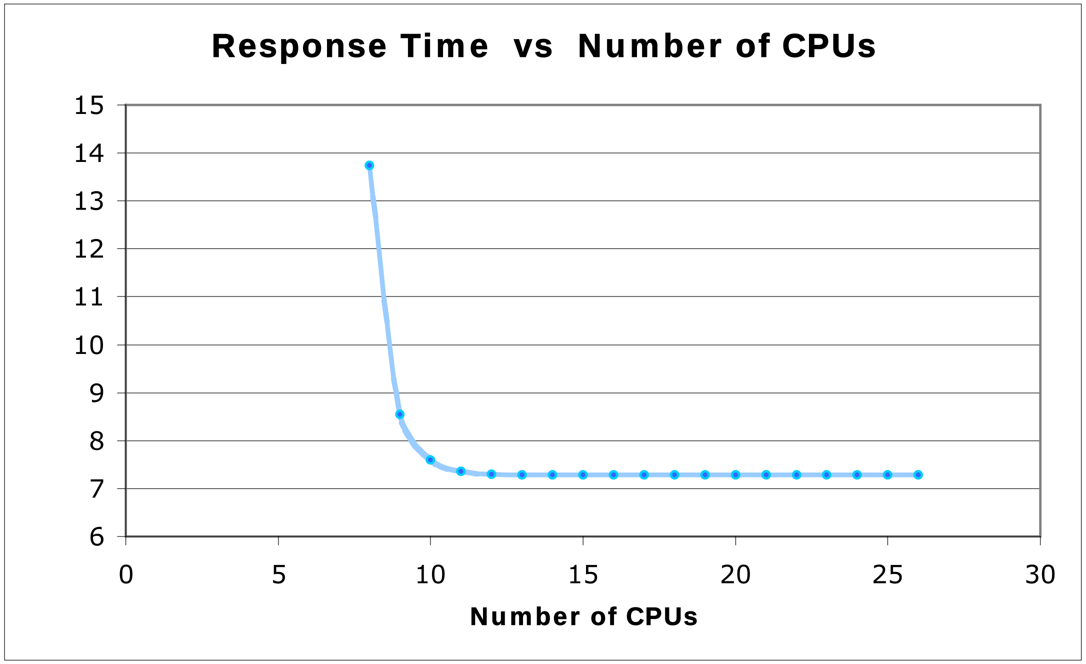
Now we can calculate the updated utilization and also the response time while changing the number of CPUs (M). We can use the formulas;

$$U = S \lambda / M$$

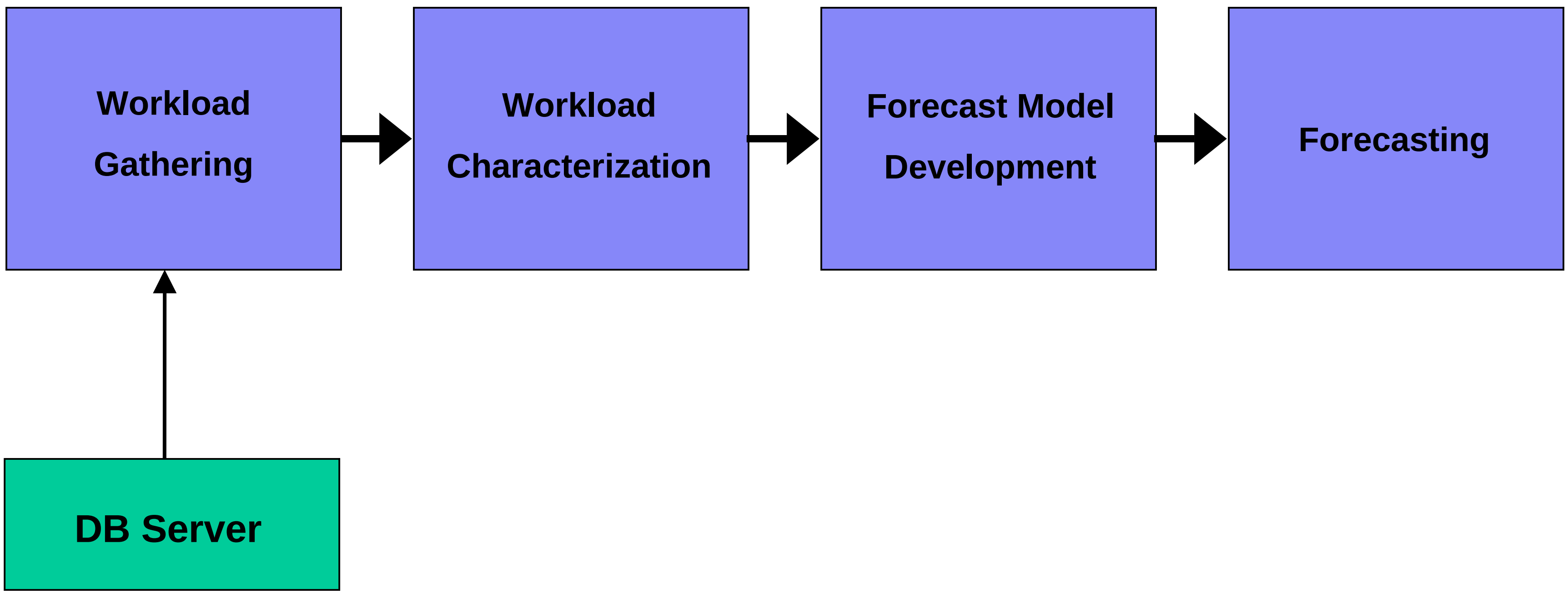
$$R = S / (1 - U^M)$$

Bob's Solution

M	% Busy	Respone Time	Queue Time	RT % Change
26	28%	7.2800	0.0000	0
25	29%	7.2800	0.0000	0.0%
24	30%	7.2800	0.0000	0.0%
23	32%	7.2800	0.0000	0.0%
22	33%	7.2800	0.0000	0.0%
21	35%	7.2800	0.0000	0.0%
20	36%	7.2800	0.0000	0.0%
19	38%	7.2800	0.0000	0.0%
18	40%	7.2800	0.0000	0.0%
17	43%	7.2800	0.0000	0.0%
16	46%	7.2800	0.0000	0.0%
15	49%	7.2801	0.0001	0.0%
14	52%	7.2808	0.0008	0.0%
13	56%	7.2839	0.0039	0.1%
12	61%	7.2981	0.0181	0.2%
11	66%	7.3585	0.0785	1.1%
10	73%	7.5977	0.3177	4.4%
9	81%	8.5471	1.2671	17.4%
8	91%	13.7424	6.4624	88.8%
7	104%	-23.0429	-30.3229	-416.5%
6	121%	-3.3232	-10.6032	-145.6%
5	146%	-1.3133	-8.5933	-118.0%
4	182%	-0.7300	-8.0100	-110.0%
3	243%	-0.5478	-7.8278	-107.5%
2	364%	-0.5943	-7.8743	-108.2%
1	728%	-1.1592	-8.4392	-115.9%



What makes a forecast more precise?

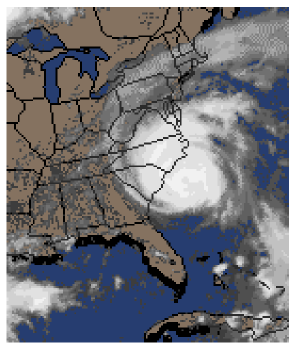


What does NOT make a forecast more precise.

- A more detailed model.
- More granular workload data.
- Sexy graphics.
- Using an inappropriate workload “peak”.

If you want to learn more, go to OraPub.com/forecast

Essential
Performance Forecasting

NYOUG - December 2005

Craig A. Shallahamer - craig@orapub.com