Performance metrics for consolidated workloads

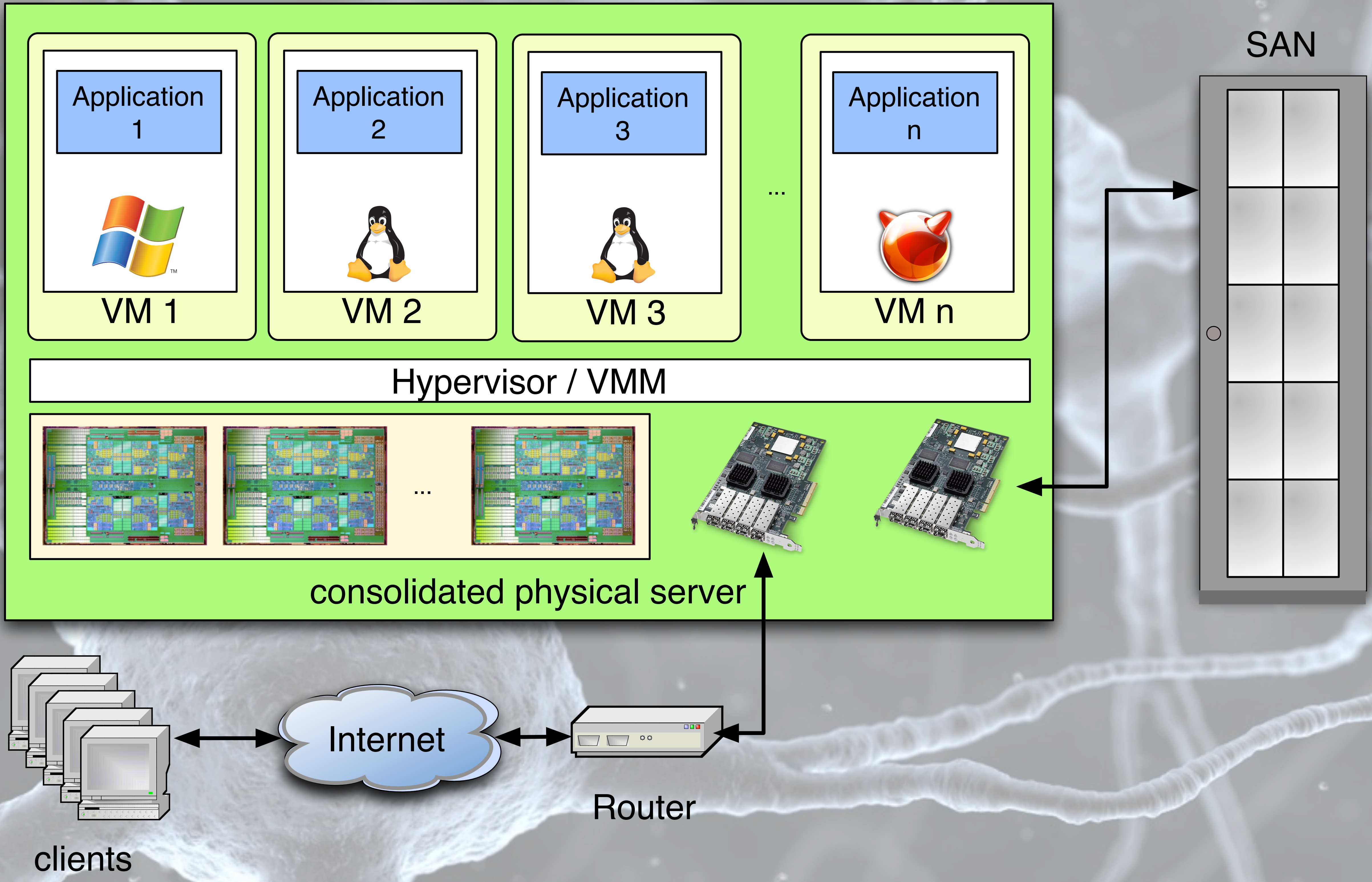
Andy Georges, Lieven Eeckhout
Computer Systems Lab
Department of Electronics and Information Systems
Ghent University

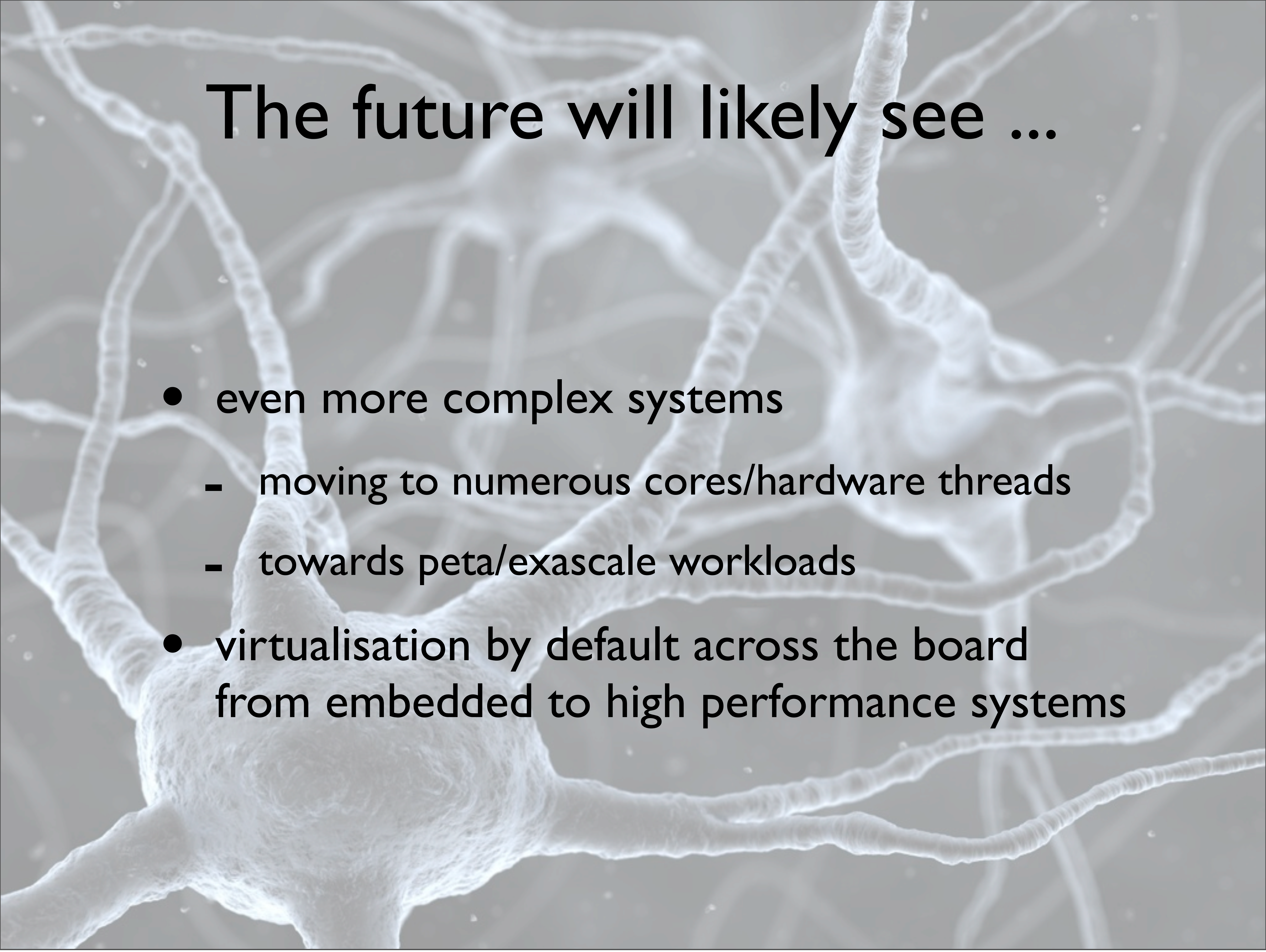
April 13, 2010

About: Me.

- Master in numerical computer sciences from Ghent University 1999
- PhD. on Java performance evaluation from Ghent University in 2008
- Post-doc for the FWO-Flanders since October 2008: performance modeling of system VMs
- Interests: performance analysis, machine learning for model building, workload characterisation, benchmarking, virtualisation, Haskell, ...

Problem statement





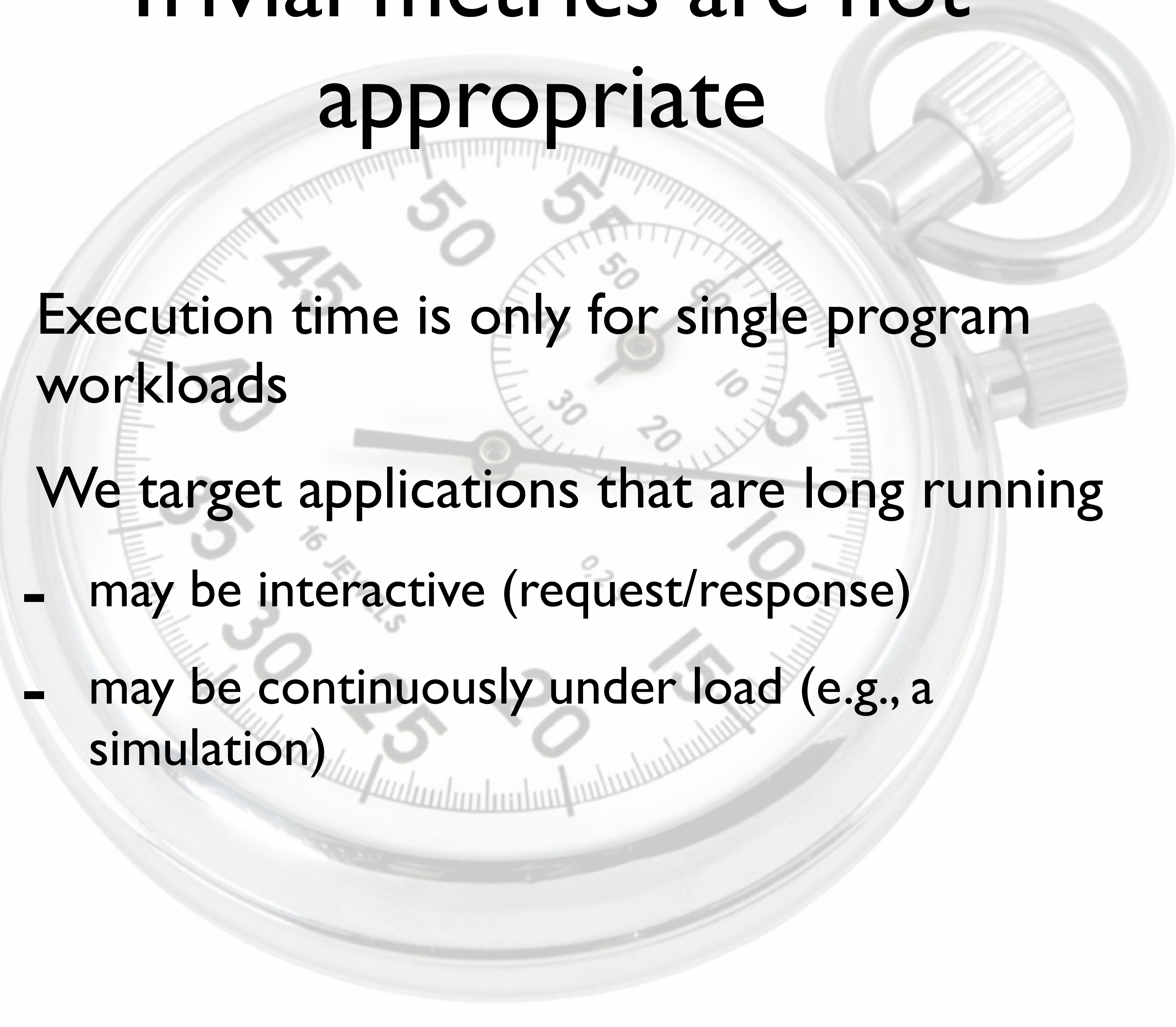
The future will likely see ...

- even more complex systems
  - moving to numerous cores/hardware threads
  - towards peta/exascale workloads
- virtualisation by default across the board from embedded to high performance systems

What do we want?

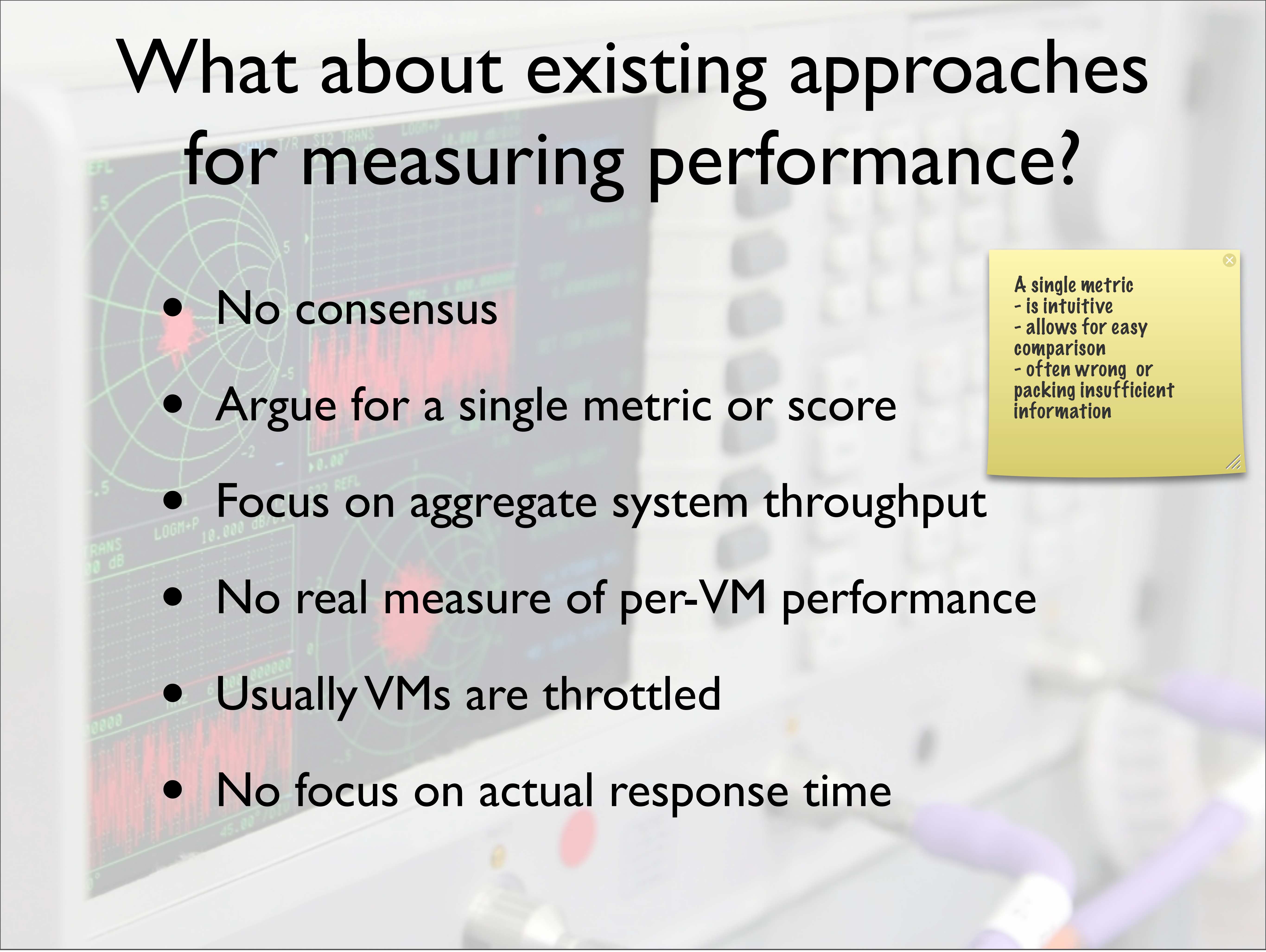
- A universal metric
- Usable in all consolidated scenario’s
- Intuitive
- Meaning at the system level
- Easy to measure

Trivial metrics are not appropriate



- Execution time is only for single program workloads
- We target applications that are long running
  - may be interactive (request/response)
  - may be continuously under load (e.g., a simulation)

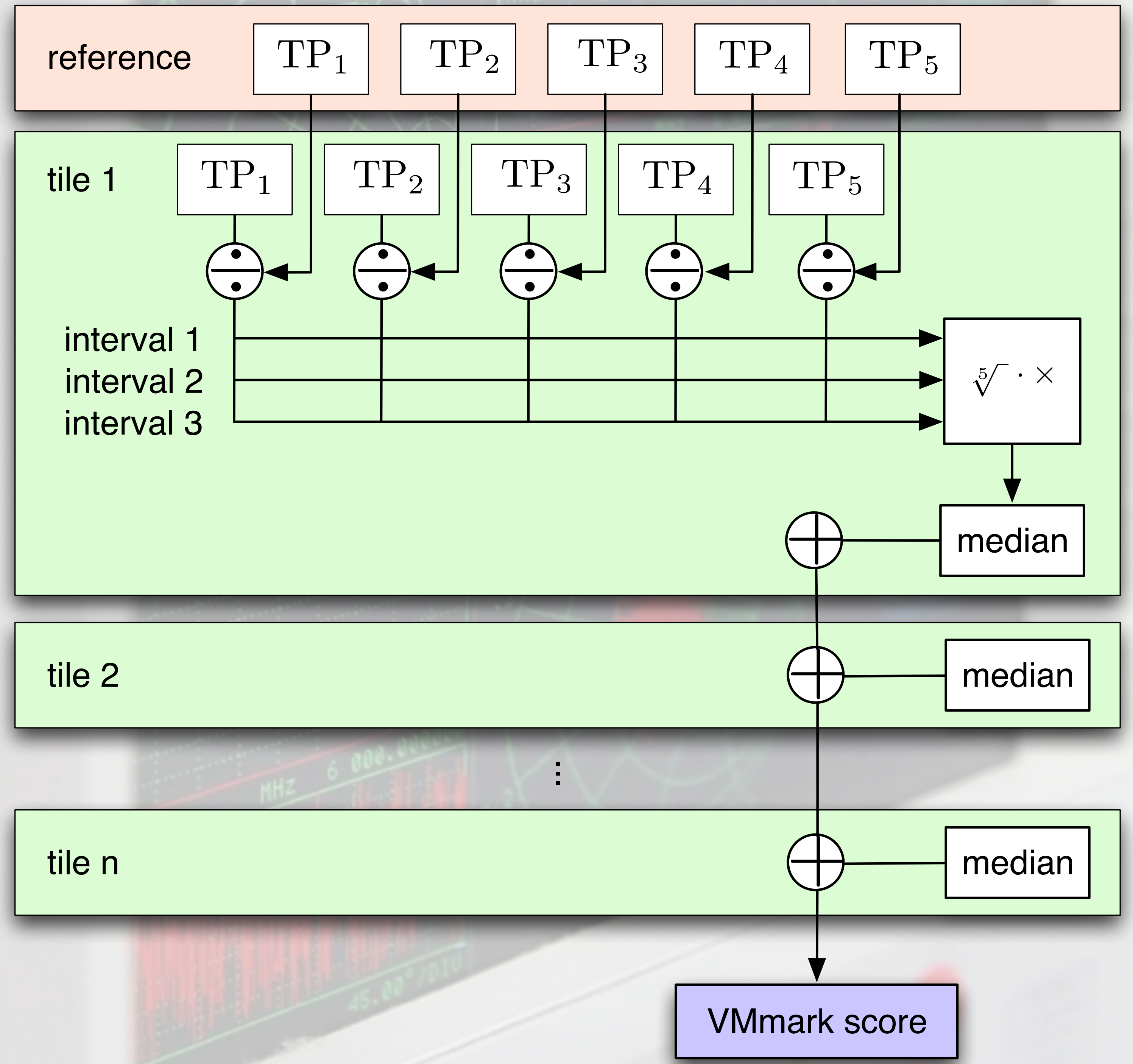
What about existing approaches for measuring performance?



- No consensus
- Argue for a single metric or score
- Focus on aggregate system throughput
- No real measure of per-VM performance
- Usually VMs are throttled
- No focus on actual response time

A single metric
- is intuitive
- allows for easy comparison
- often wrong or packing insufficient information

VMmark



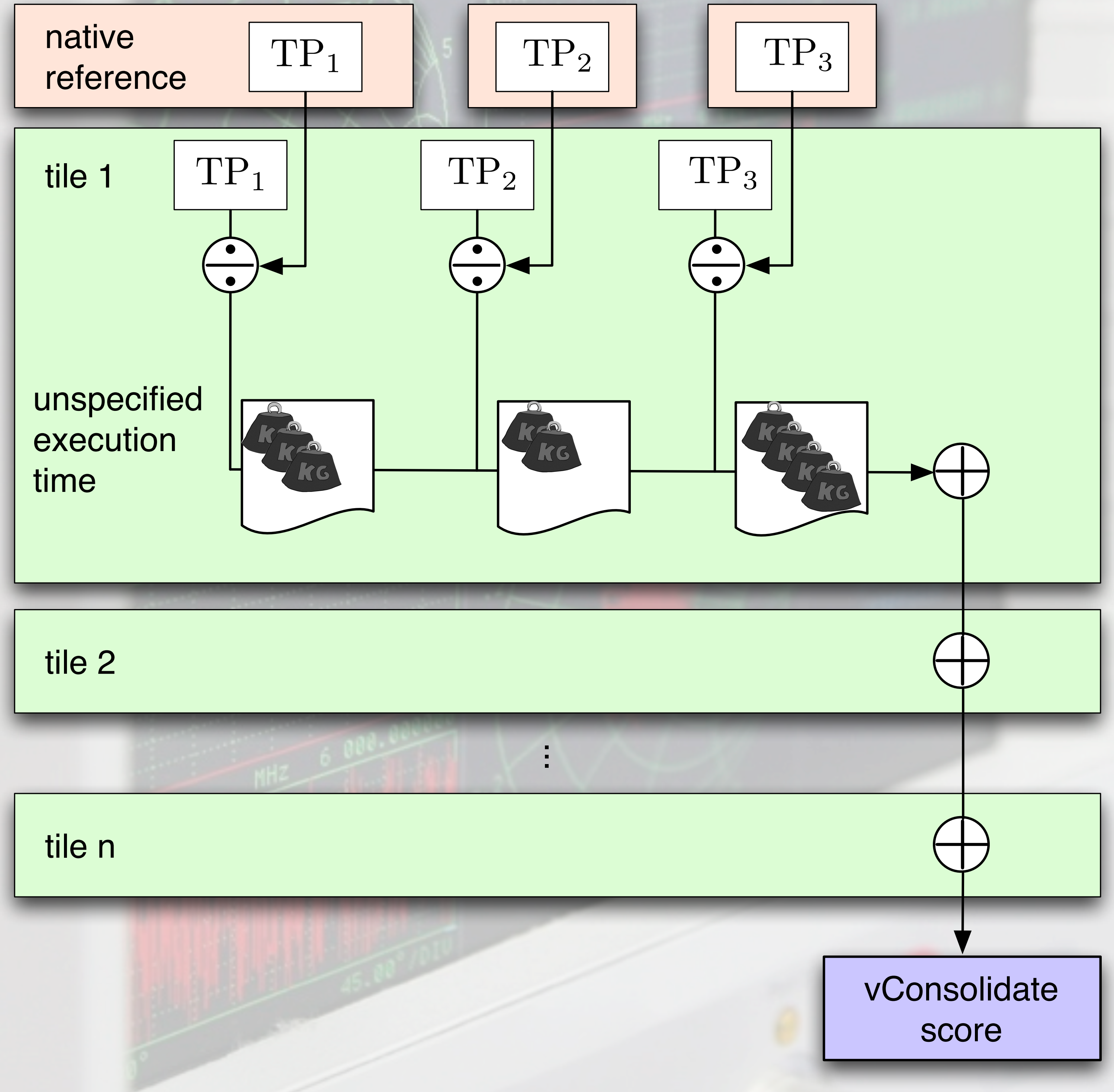
State something on the use of means.

geomean is only to be used for dependent values, such as the interest rates in a bank of paycheck raises over multiple years.

- Benchmark
  - MS excha
  - SPECjbb2000
  - SPECweb2005
  - Swingbench
  - dbench
- sum over tiles of median over intervals of geomean over benchmarks

VMmark: A Scalable Benchmark for Virtualized Systems,
V. Makhija, B. Herndon, P. Smith, L. Roderick, E. Zamost, J. Anderson, Technical Report VMware-TR-2006-002, 2006

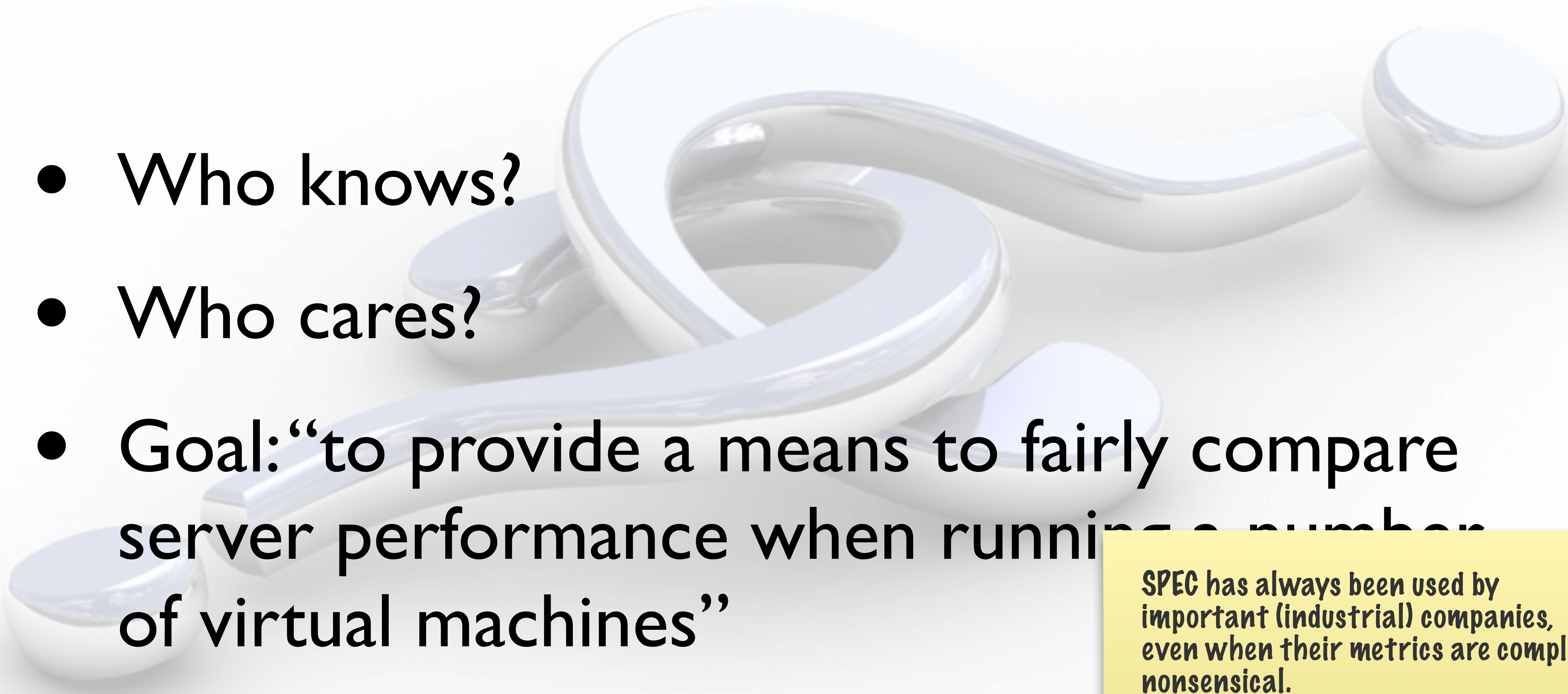
vConsolidate



- Benchmarks
  - webserver
  - mailserver
  - db server
- sum over tiles of weighted sum of benchmarks per tile

Redefining Server Performance Characterization for Virtualization Benchmarking, J.P. Casazza, M. Greenfield, K. Shi, Intel Tech. Journal, Vol 10(3), 2006

SPECvirt



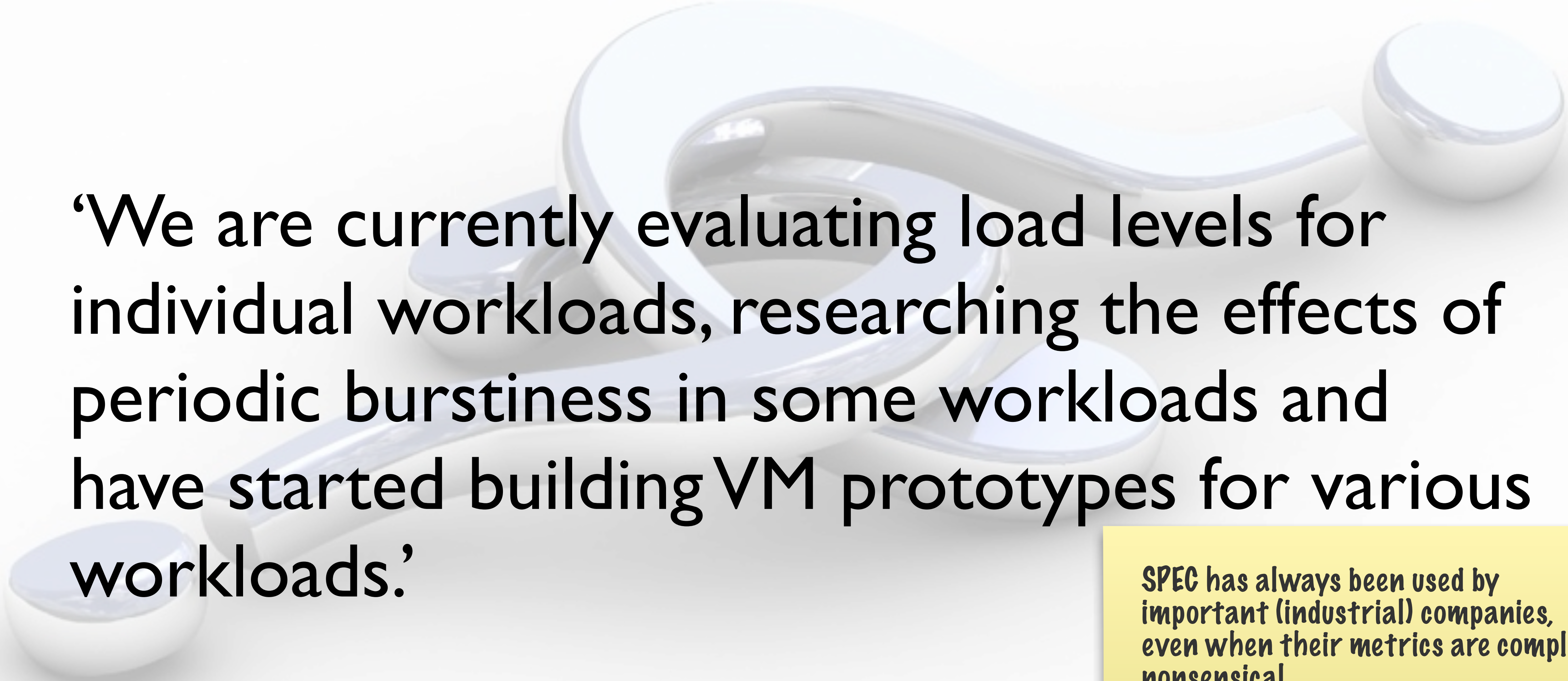
- Who knows?
- Who cares?
- Goal: "to provide a means to fairly compare server performance when running a number of virtual machines"

SPEC has always been used by important (industrial) companies, even when their metrics are complete nonsensical.

So, we should care at some point.

Preferably, we correct them before they come up with something that is hardly usable or even plain wrong. Think SPECjvm98 best and worst runtimes, think geomeans, ...

SPECvirt



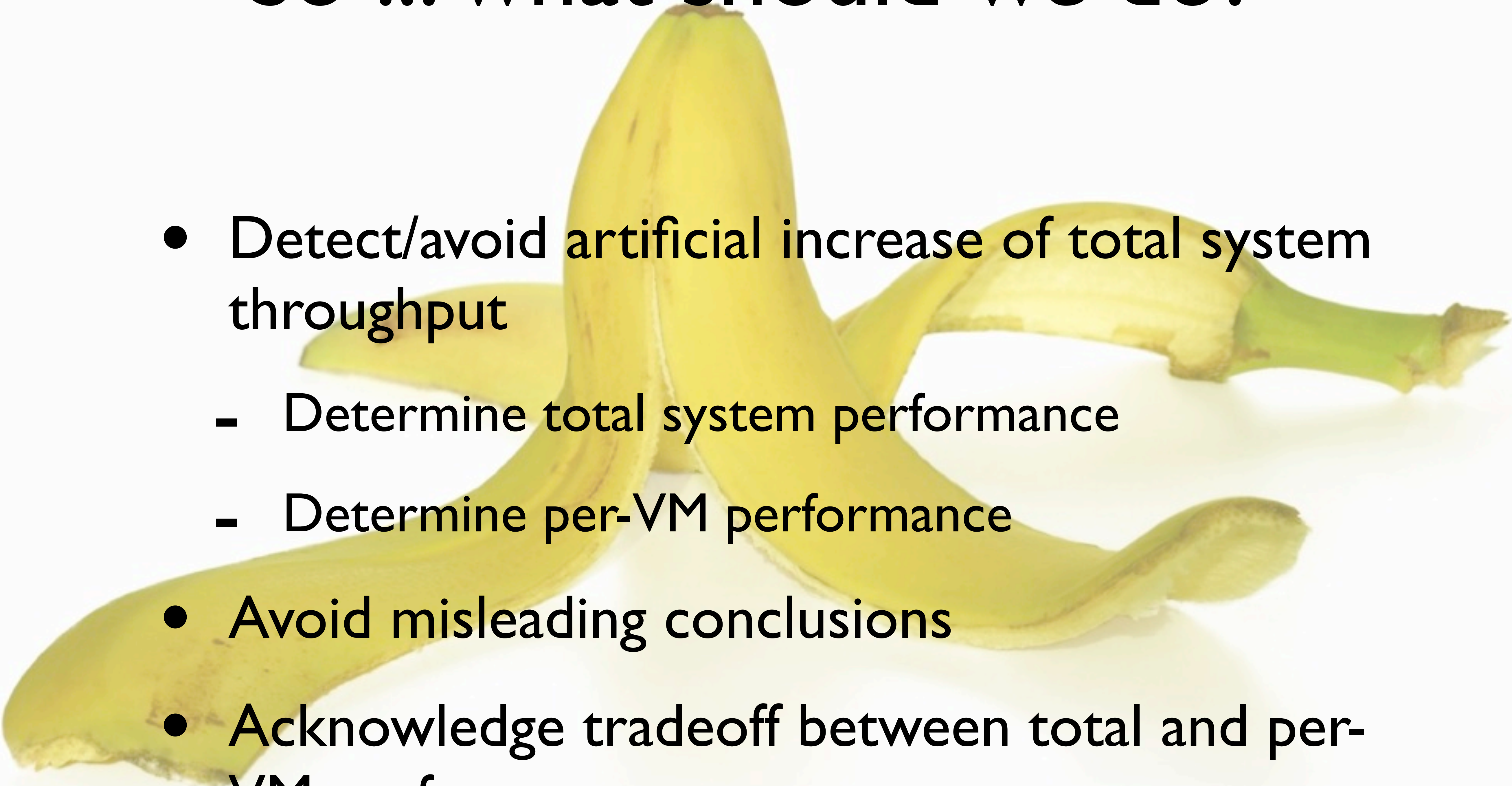
‘We are currently evaluating load levels for individual workloads, researching the effects of periodic burstiness in some workloads and have started building VM prototypes for various workloads.’

SPEC has always been used by important (industrial) companies, even when their metrics are complete nonsensical.

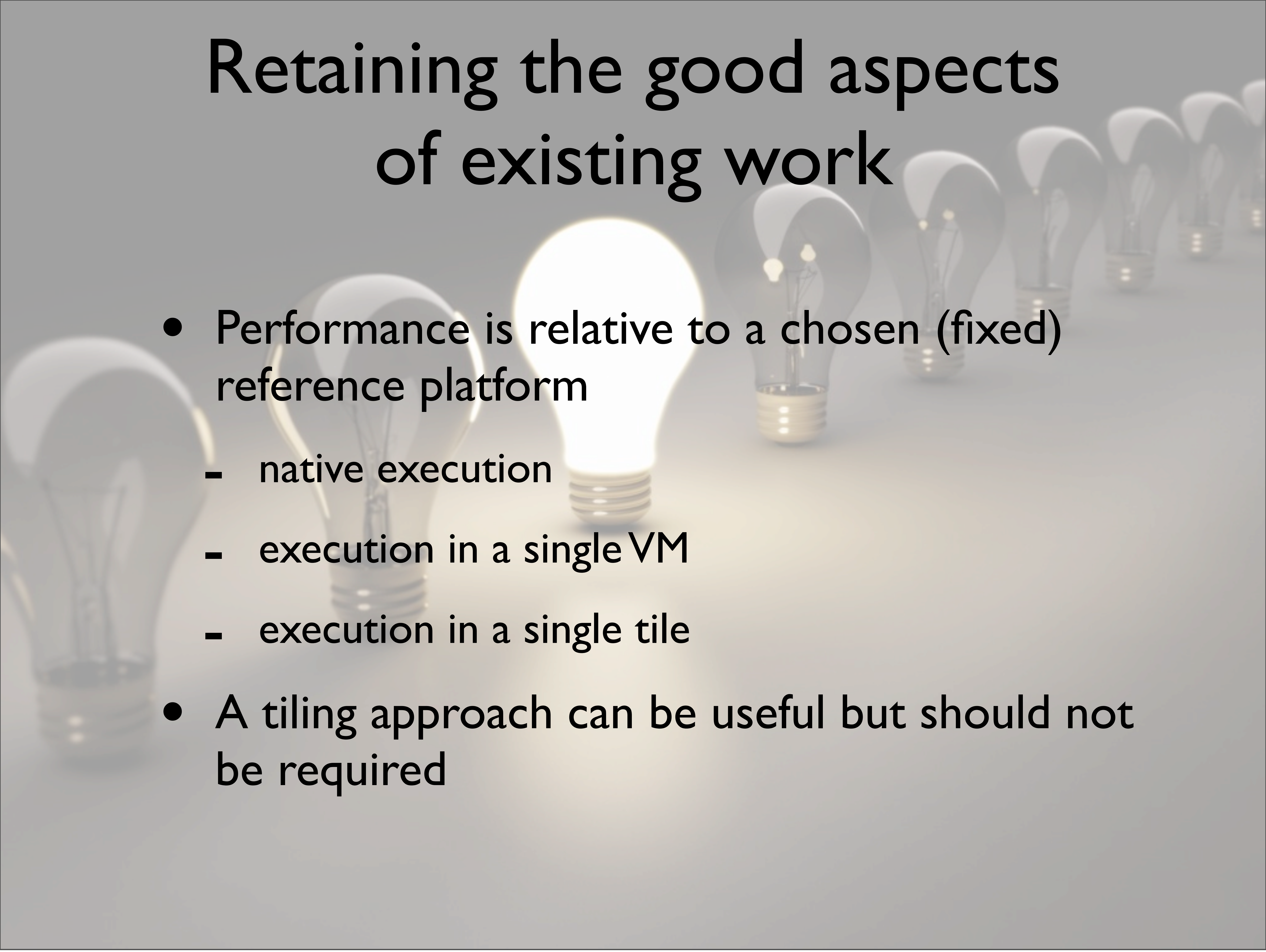
So, we should care at some point.

Preferably, we correct them before they come up with something that is hardly usable or even plain wrong. Think SPECjvm98 best and worst runtimes, think geomeans, ...

So ... what should we do?



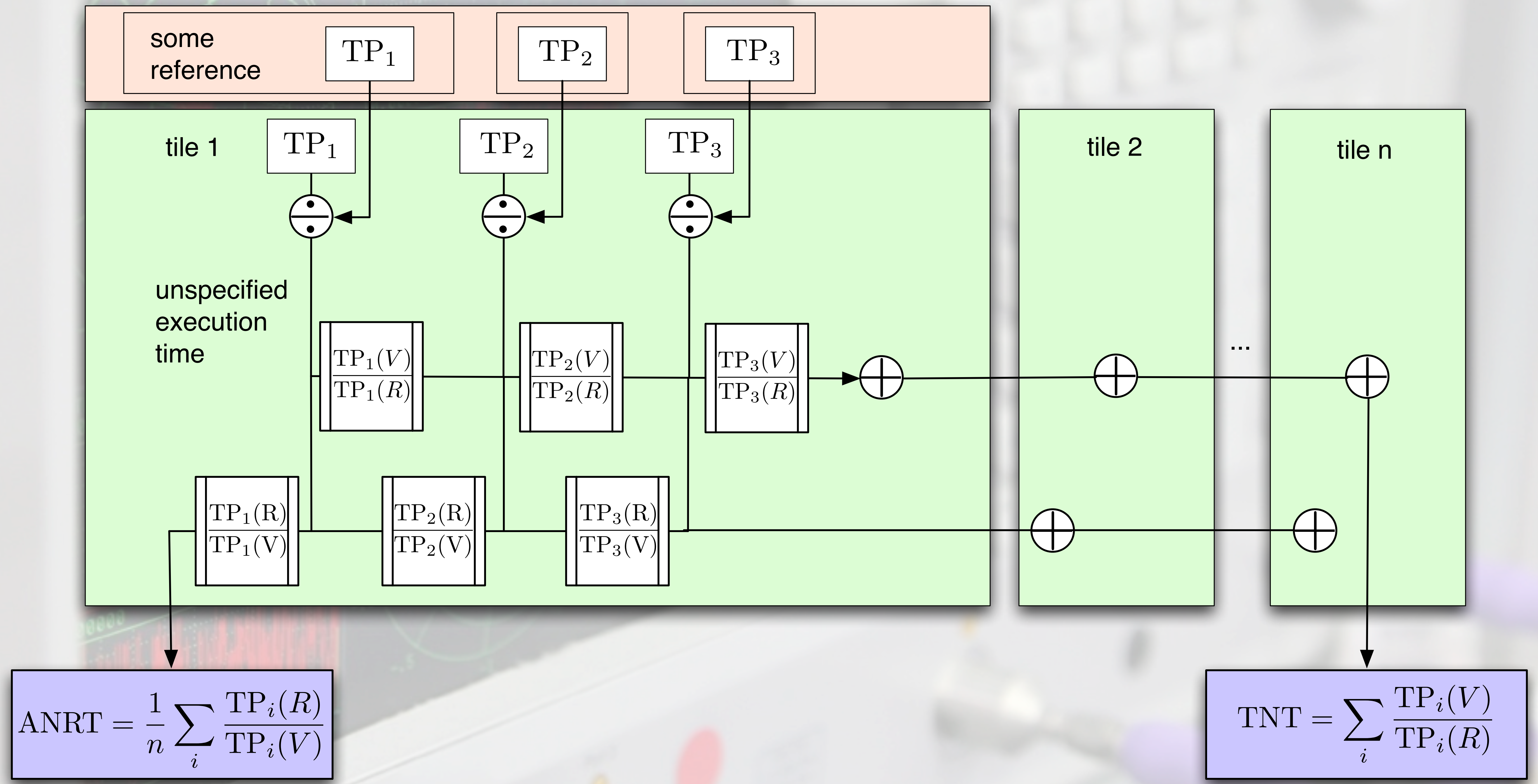
- Detect/avoid artificial increase of total system throughput
  - Determine total system performance
  - Determine per-VM performance
- Avoid misleading conclusions
- Acknowledge tradeoff between total and per-VM performance



Retaining the good aspects of existing work

- Performance is relative to a chosen (fixed) reference platform
  - native execution
  - execution in a single VM
  - execution in a single tile
- A tiling approach can be useful but should not be required

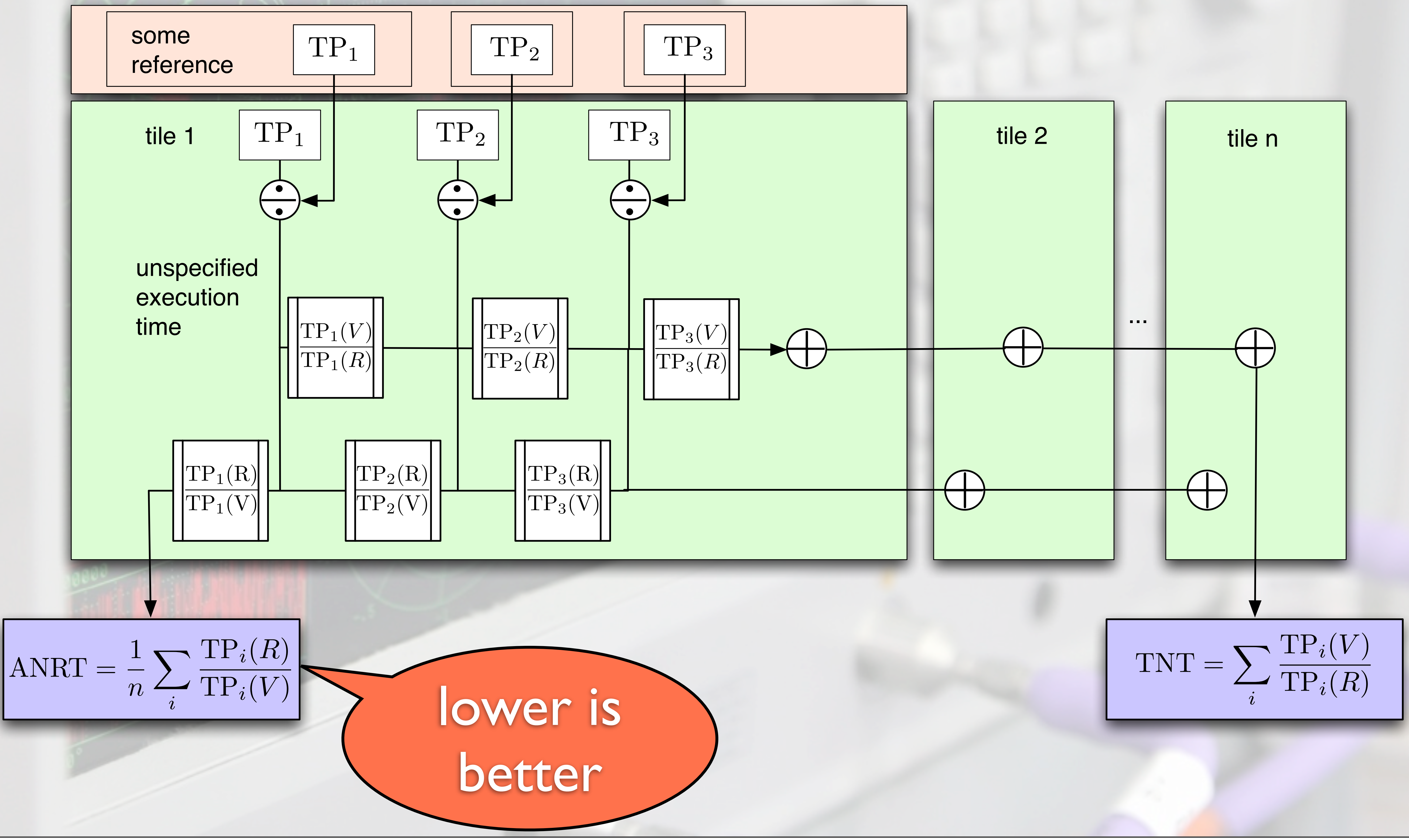
TNT and ANRT


some reference
TP_1
TP_2
TP_3
tile 1
TP_1
TP_2
TP_3
unspecified execution time
$\frac{\mathrm{TP}_1(V)}{\mathrm{TP}_1(R)}$
$\frac{\mathrm{TP}_2(V)}{\mathrm{TP}_2(R)}$
$\frac{\mathrm{TP}_3(V)}{\mathrm{TP}_3(R)}$
$\frac{\mathrm{TP}_1(\mathrm{R})}{\mathrm{TP}_1(\mathrm{V})}$
$\frac{\mathrm{TP}_2(\mathrm{R})}{\mathrm{TP}_2(\mathrm{V})}$
$\frac{\mathrm{TP}_3(\mathrm{R})}{\mathrm{TP}_3(\mathrm{V})}$
tile 2
...
tile n

$$\mathrm{ANRT} = \frac{1}{n}\sum_i \frac{\mathrm{TP}_i(R)}{\mathrm{TP}_i(V)}$$

$$\mathrm{TNT} = \sum_i \frac{\mathrm{TP}_i(V)}{\mathrm{TP}_i(R)}$$

TNT and ANRT

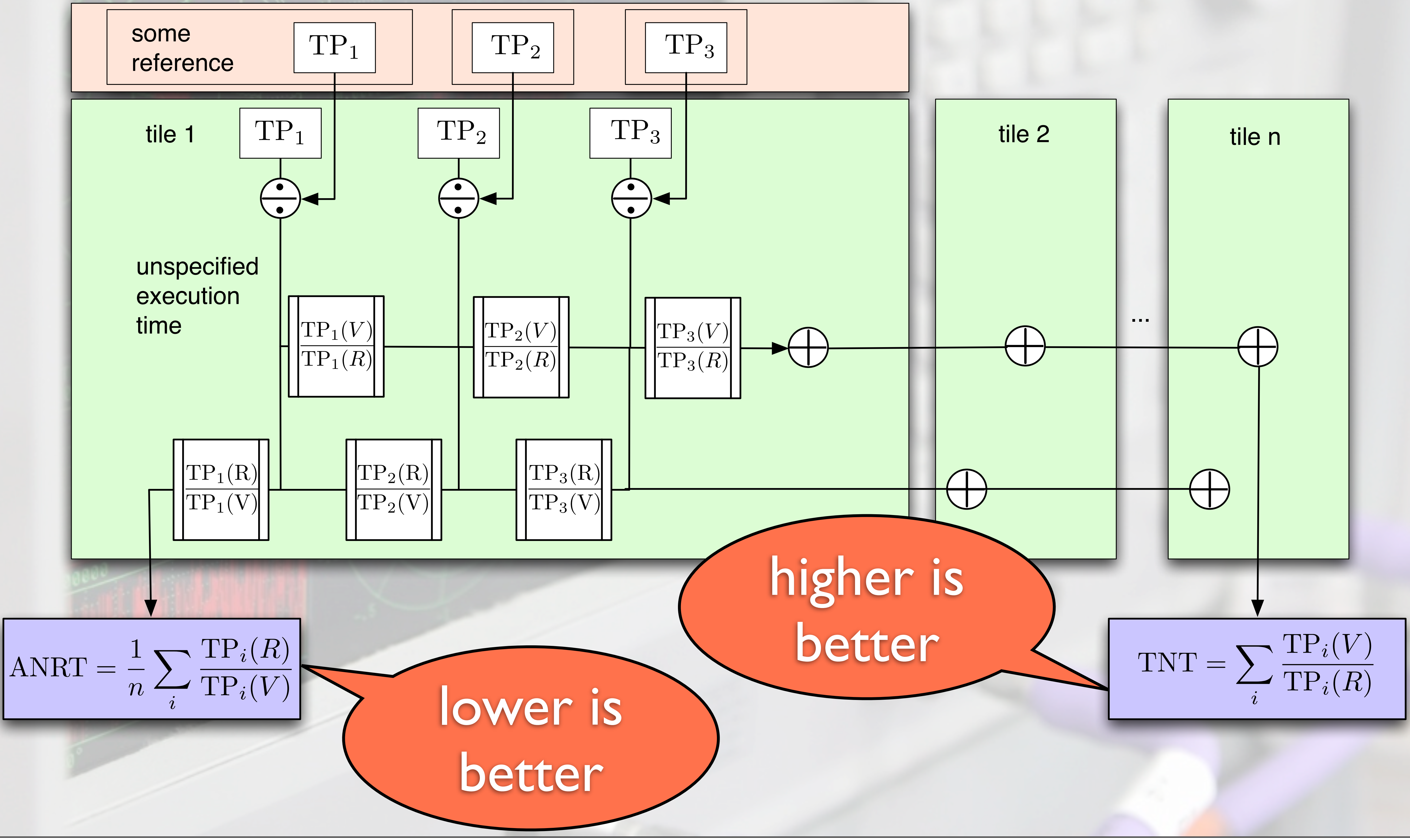

some reference
TP_1
TP_2
TP_3
tile 1
TP_1
TP_2
TP_3
unspecified execution time
$\frac{\mathrm{TP}_1(V)}{\mathrm{TP}_1(R)}$
$\frac{\mathrm{TP}_2(V)}{\mathrm{TP}_2(R)}$
$\frac{\mathrm{TP}_3(V)}{\mathrm{TP}_3(R)}$
$\frac{\mathrm{TP}_1(\mathrm{R})}{\mathrm{TP}_1(\mathrm{V})}$
$\frac{\mathrm{TP}_2(\mathrm{R})}{\mathrm{TP}_2(\mathrm{V})}$
$\frac{\mathrm{TP}_3(\mathrm{R})}{\mathrm{TP}_3(\mathrm{V})}$
tile 2
...
tile n

$$\mathrm{ANRT} = \frac{1}{n}\sum_i \frac{\mathrm{TP}_i(R)}{\mathrm{TP}_i(V)}$$

lower is better

$$\mathrm{TNT} = \sum_i \frac{\mathrm{TP}_i(V)}{\mathrm{TP}_i(R)}$$

TNT and ANRT


some reference
TP_1
TP_2
TP_3
tile 1
TP_1
TP_2
TP_3
unspecified execution time
$\frac{TP_1(V)}{TP_1(R)}$
$\frac{TP_2(V)}{TP_2(R)}$
$\frac{TP_3(V)}{TP_3(R)}$
$\frac{TP_1(R)}{TP_1(V)}$
$\frac{TP_2(R)}{TP_2(V)}$
$\frac{TP_3(R)}{TP_3(V)}$
tile 2
...
tile n

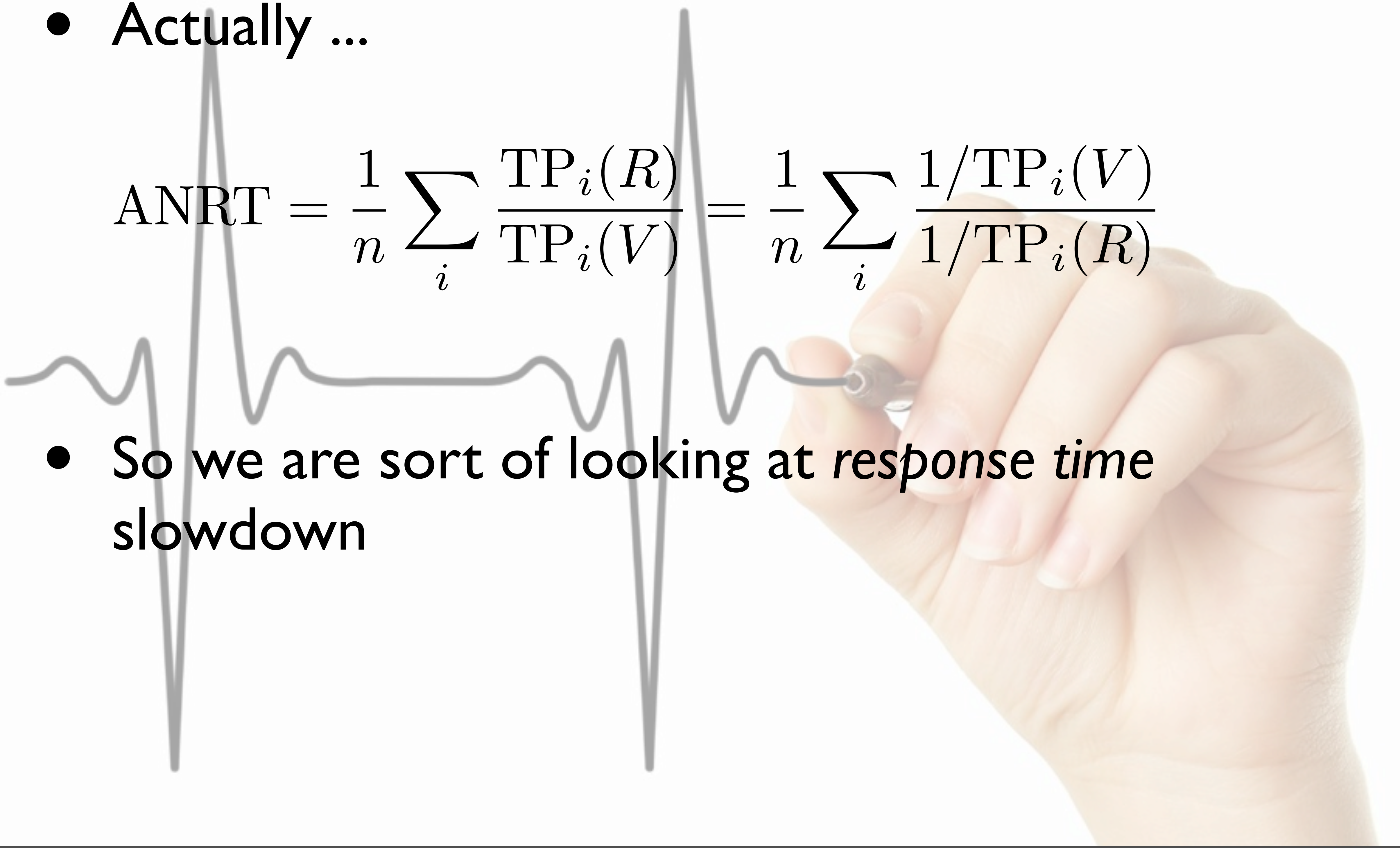
$$\mathrm{ANRT} = \frac{1}{n}\sum_i \frac{\mathrm{TP}_i(R)}{\mathrm{TP}_i(V)}$$

lower is better

$$\mathrm{TNT} = \sum_i \frac{\mathrm{TP}_i(V)}{\mathrm{TP}_i(R)}$$

higher is better

Proper use of mean, eh?



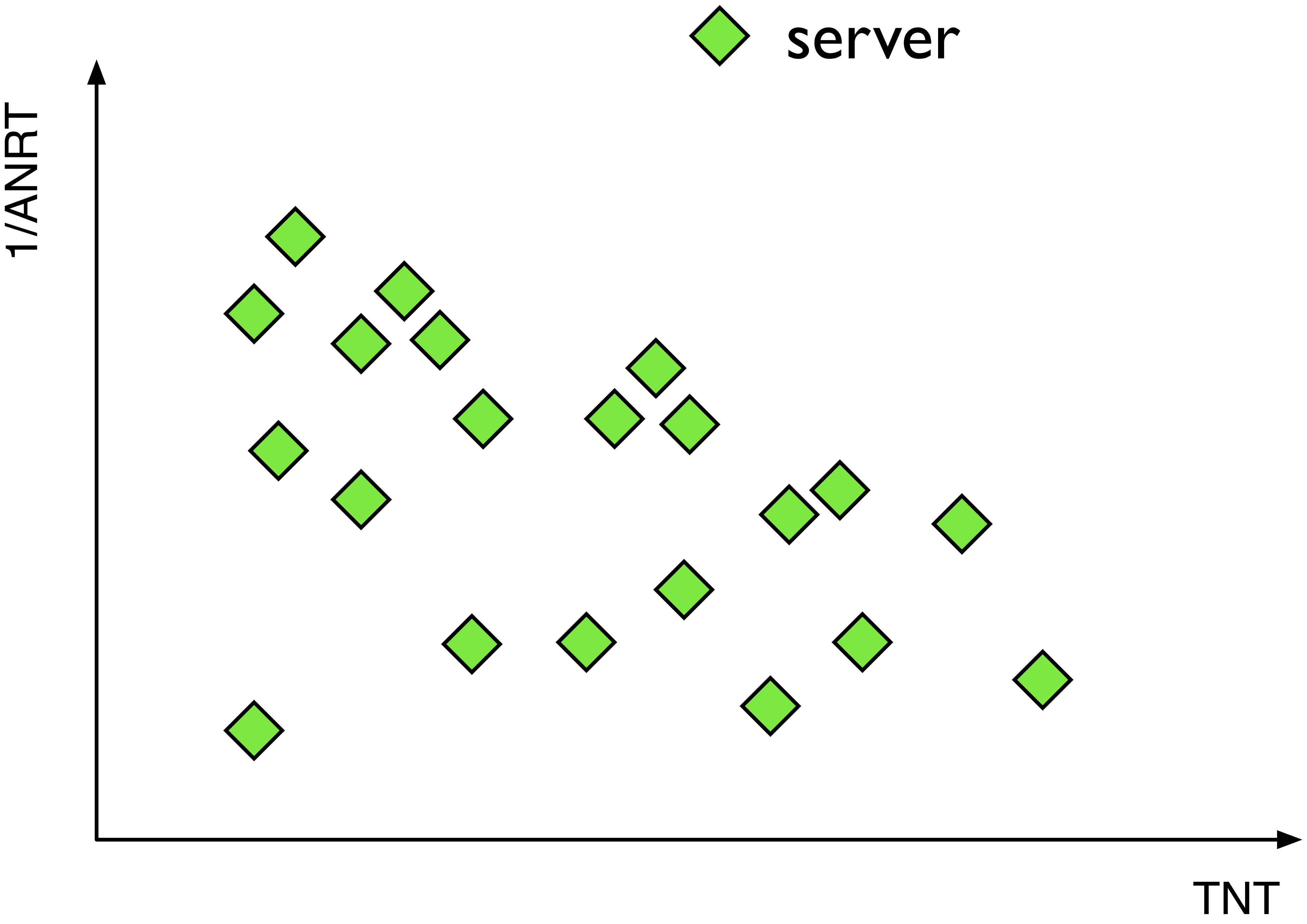
- Actually ...

$$\mathrm{ANRT} = \frac{1}{n}\sum_i \frac{\mathrm{TP}_i(R)}{\mathrm{TP}_i(V)} = \frac{1}{n}\sum_i \frac{1/\mathrm{TP}_i(V)}{1/\mathrm{TP}_i(R)}$$

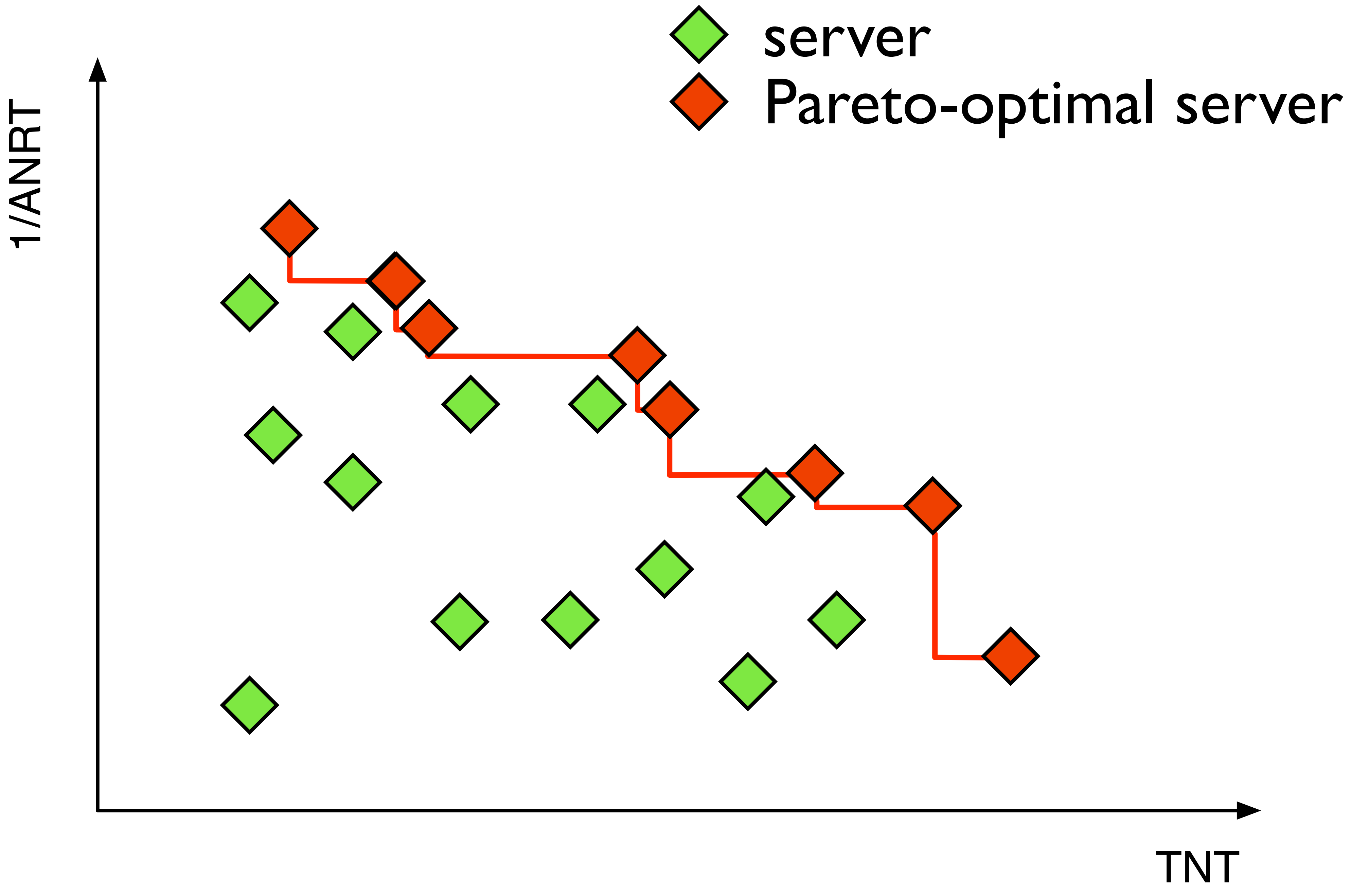
- So we are sort of looking at *response time* slowdown

Pareto curves

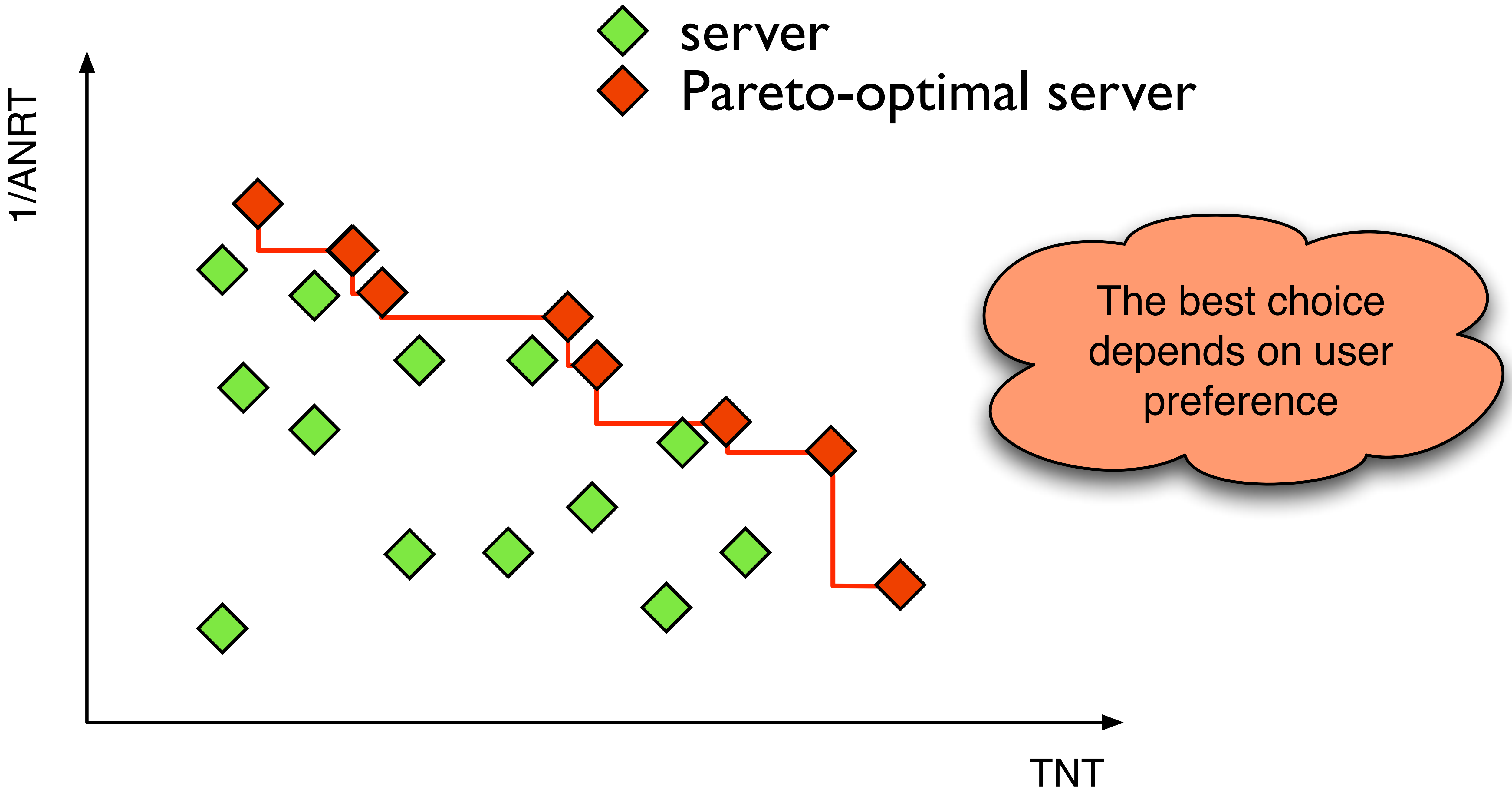
Pareto curves



Pareto curves

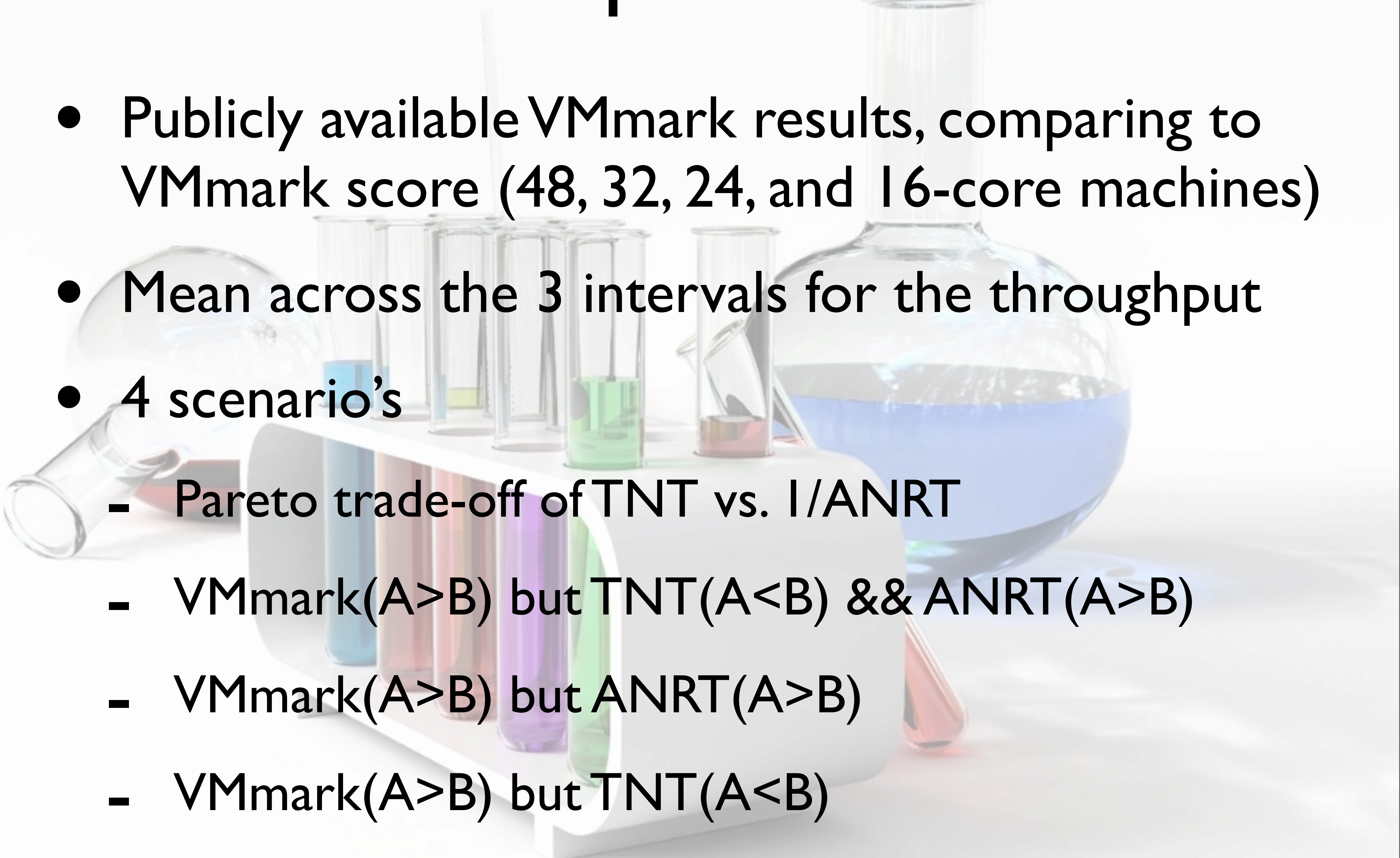


Pareto curves



The best choice depends on user preference

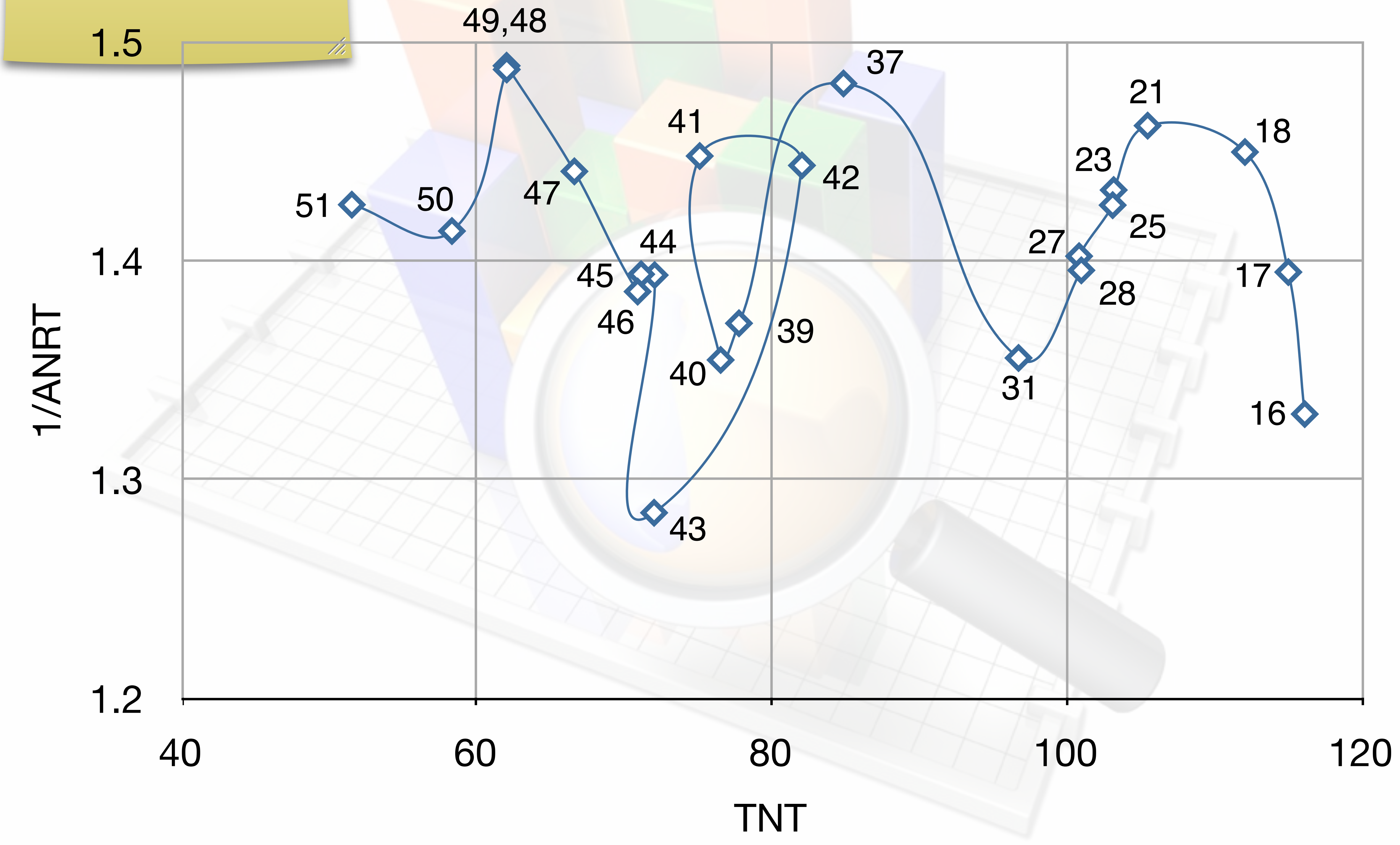
Some experiments



- Publicly available VMmark results, comparing to VMmark score (48, 32, 24, and 16-core machines)
- Mean across the 3 intervals for the throughput
- 4 scenario's
  - Pareto trade-off of TNT vs. 1/ANRT
  - VMmark(A>B) but TNT(A<B) && ANRT(A>B)
  - VMmark(A>B) but ANRT(A>B)
  - VMmark(A>B) but TNT(A<B)

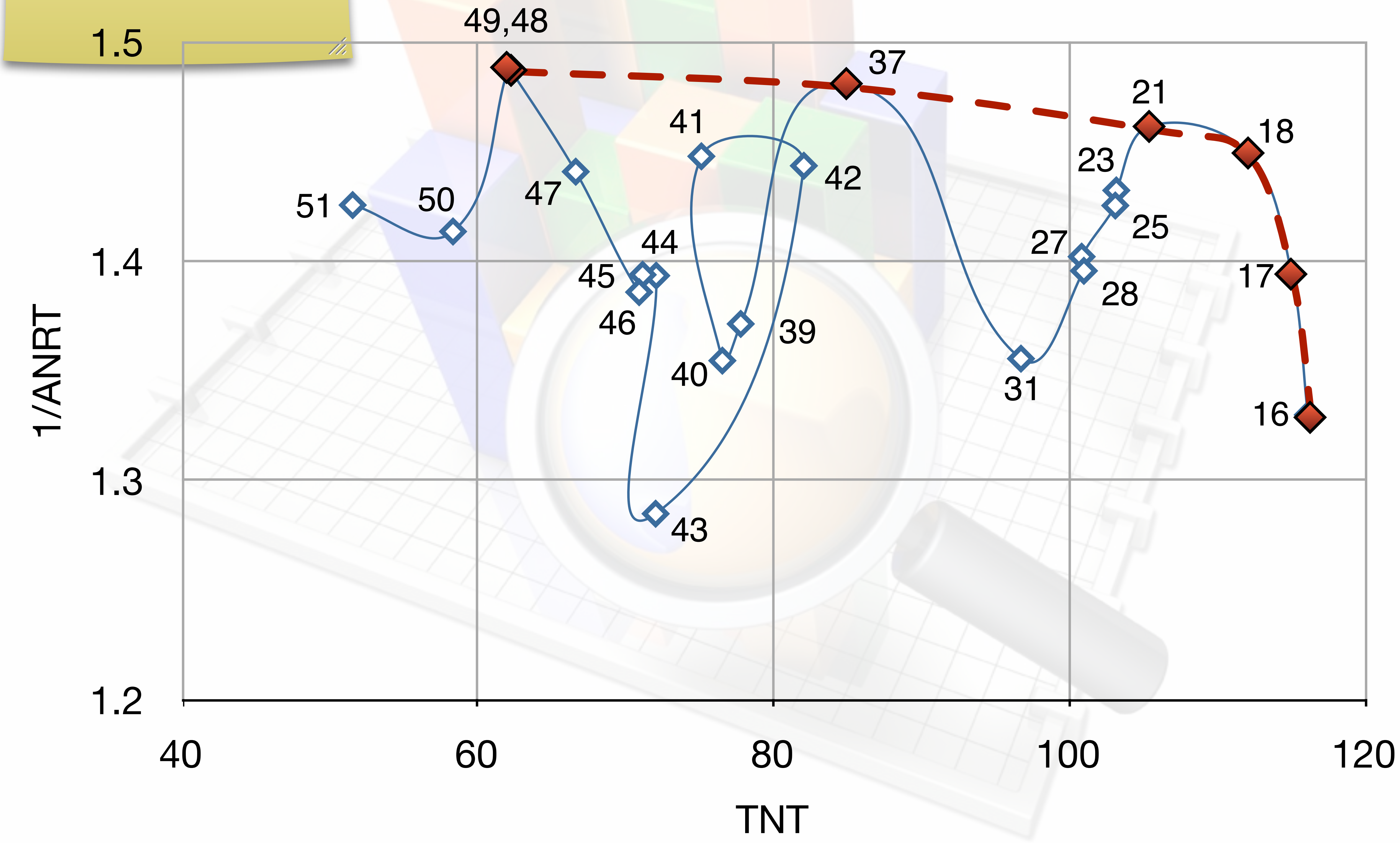
- pareto curve (red)
- inverting for TNT (brown
- inverting for ANRT (green)

16-core systems



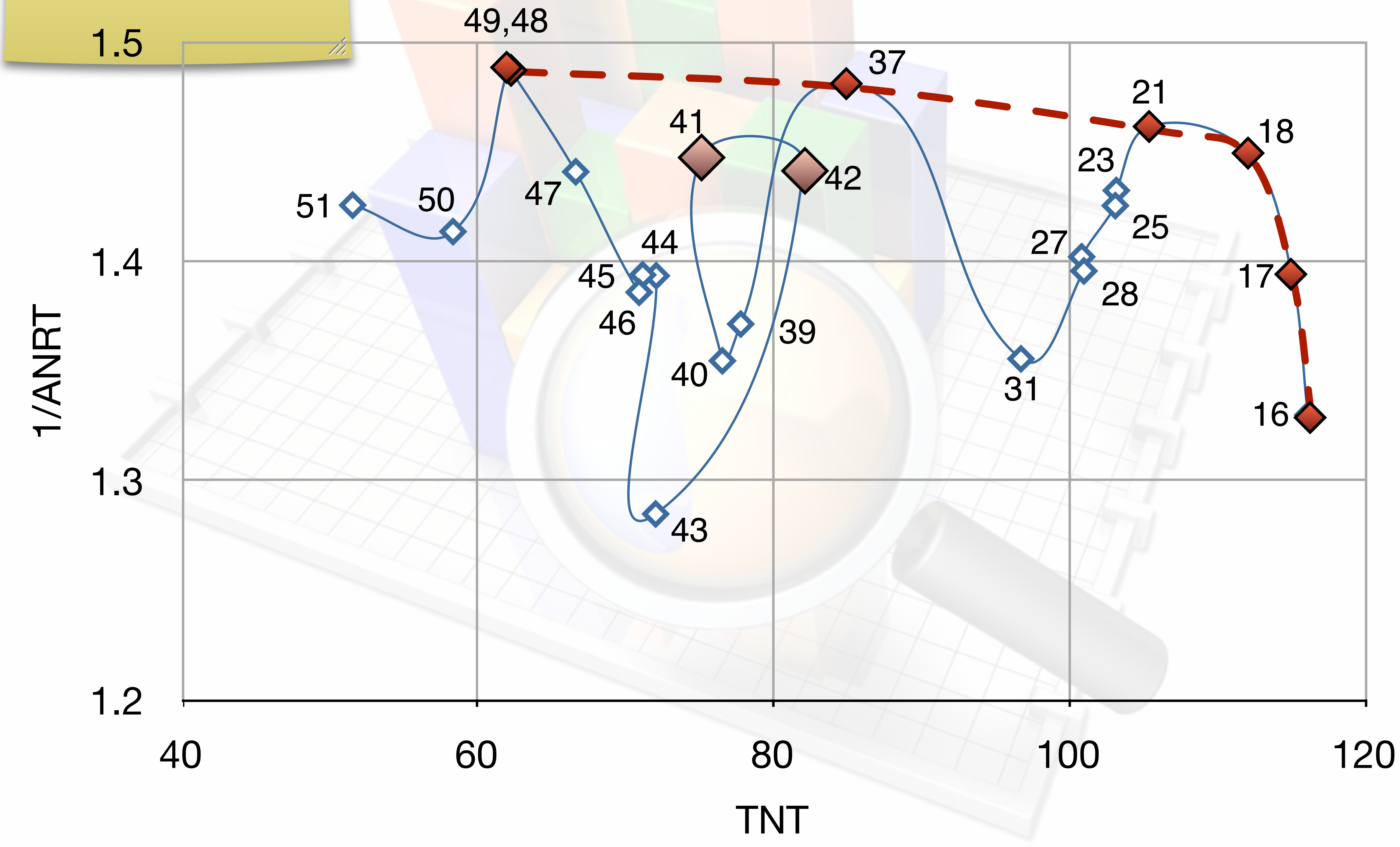
- pareto curve (red)
- inverting for TNT (brown
- inverting for ANRT (green)

16-core systems



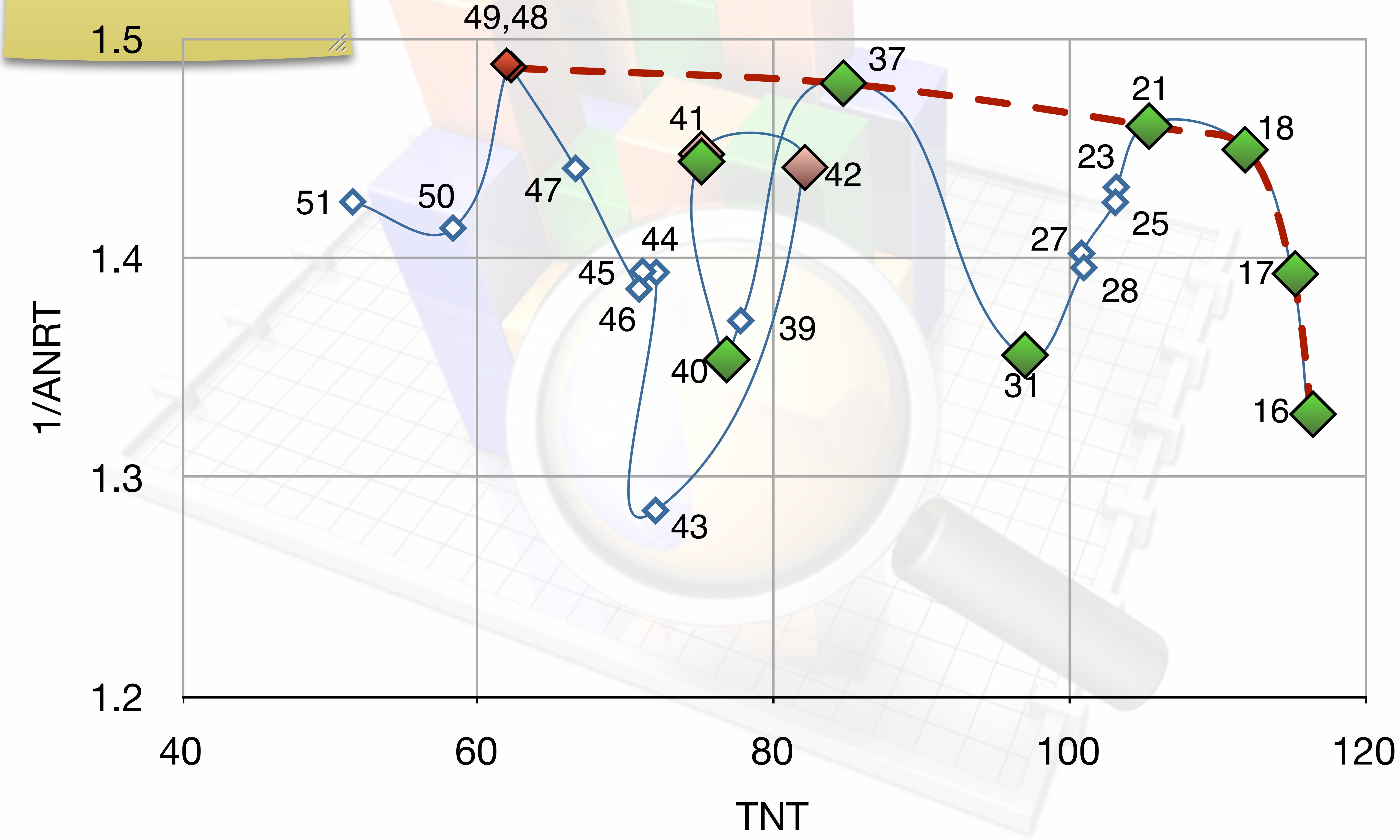
- pareto curve (red)
- inverting for TNT (brown
- inverting for ANRT (green)

16-core systems

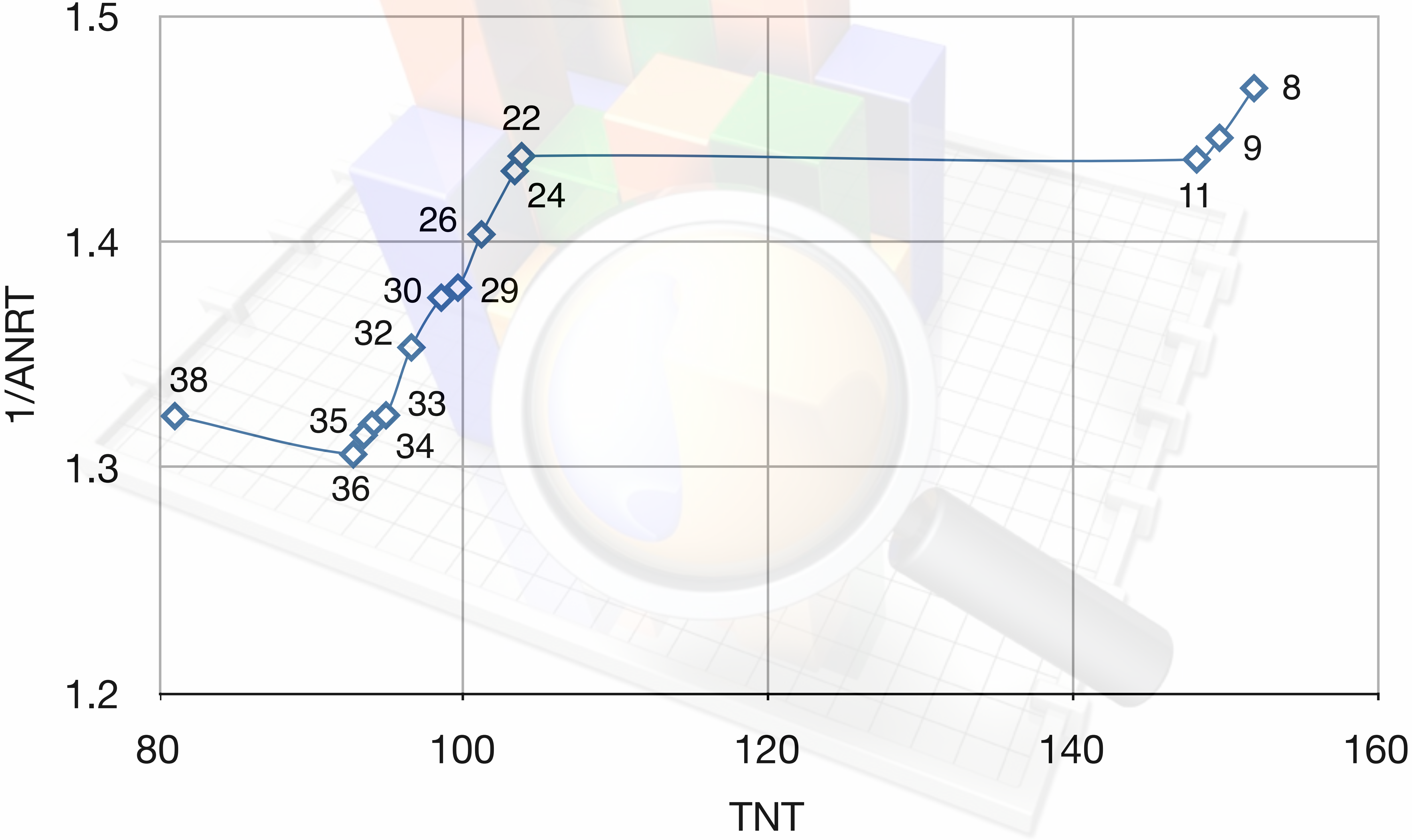


- pareto curve (red)
- inverting for TNT (brown
- inverting for ANRT (green)

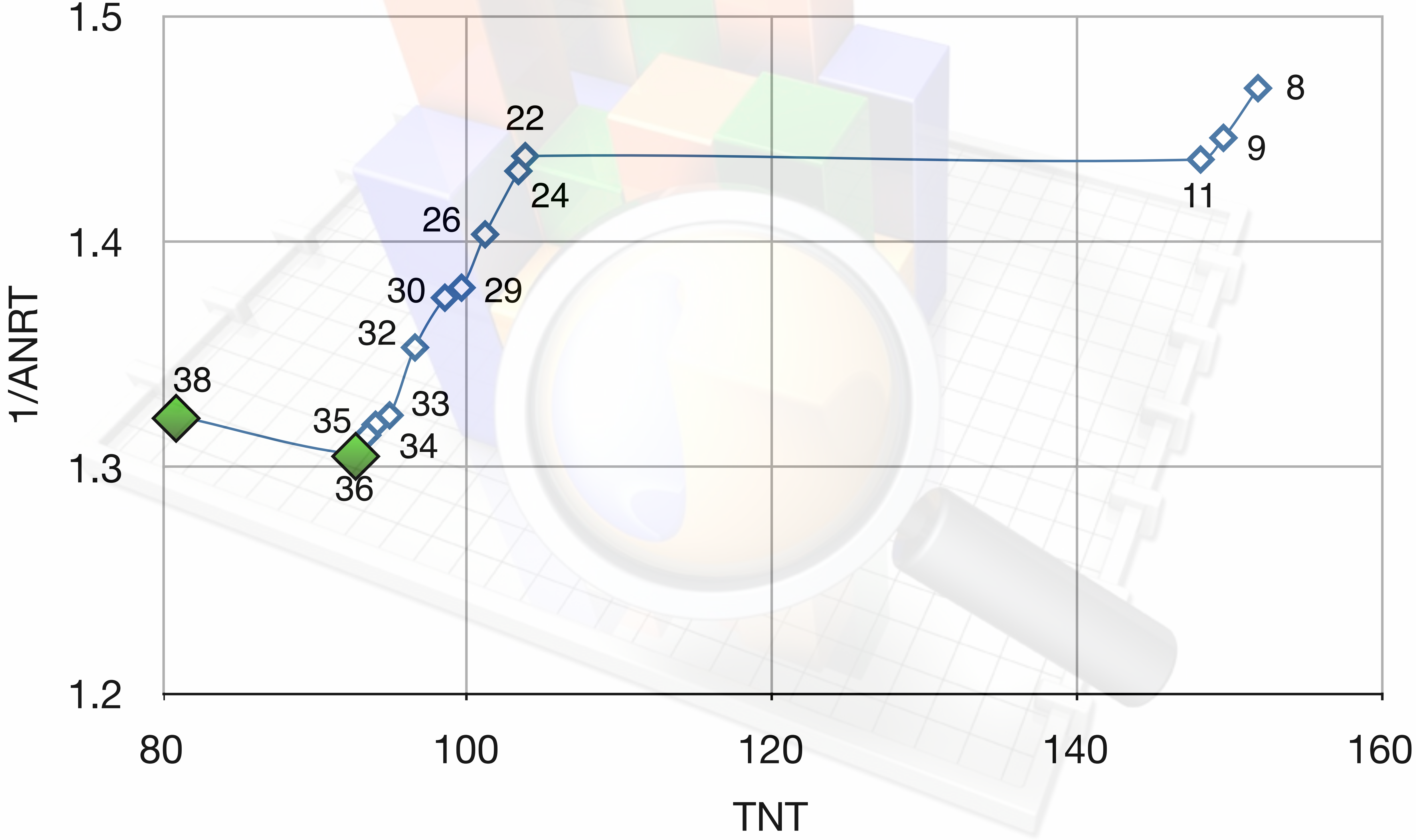
16-core systems



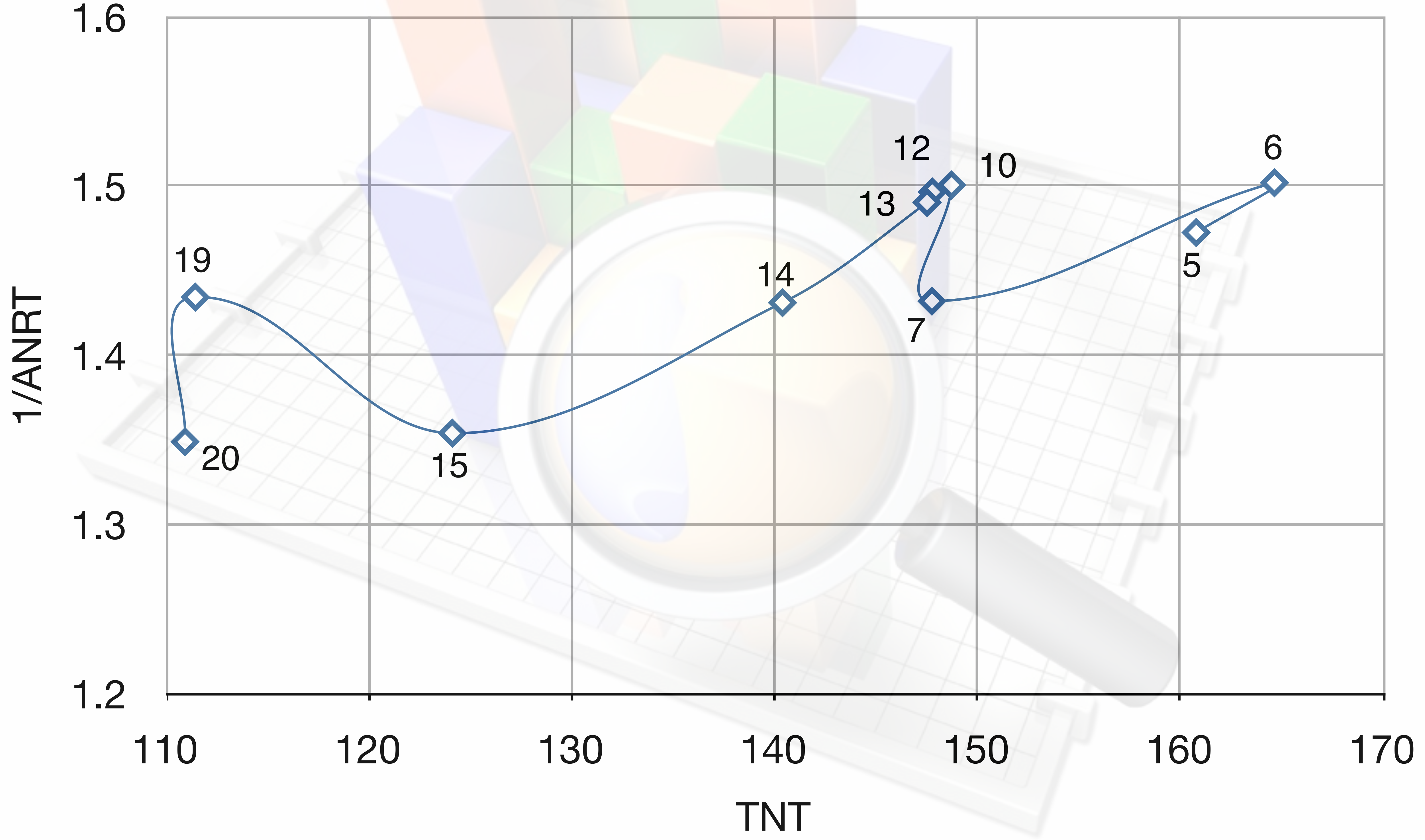
24-core systems



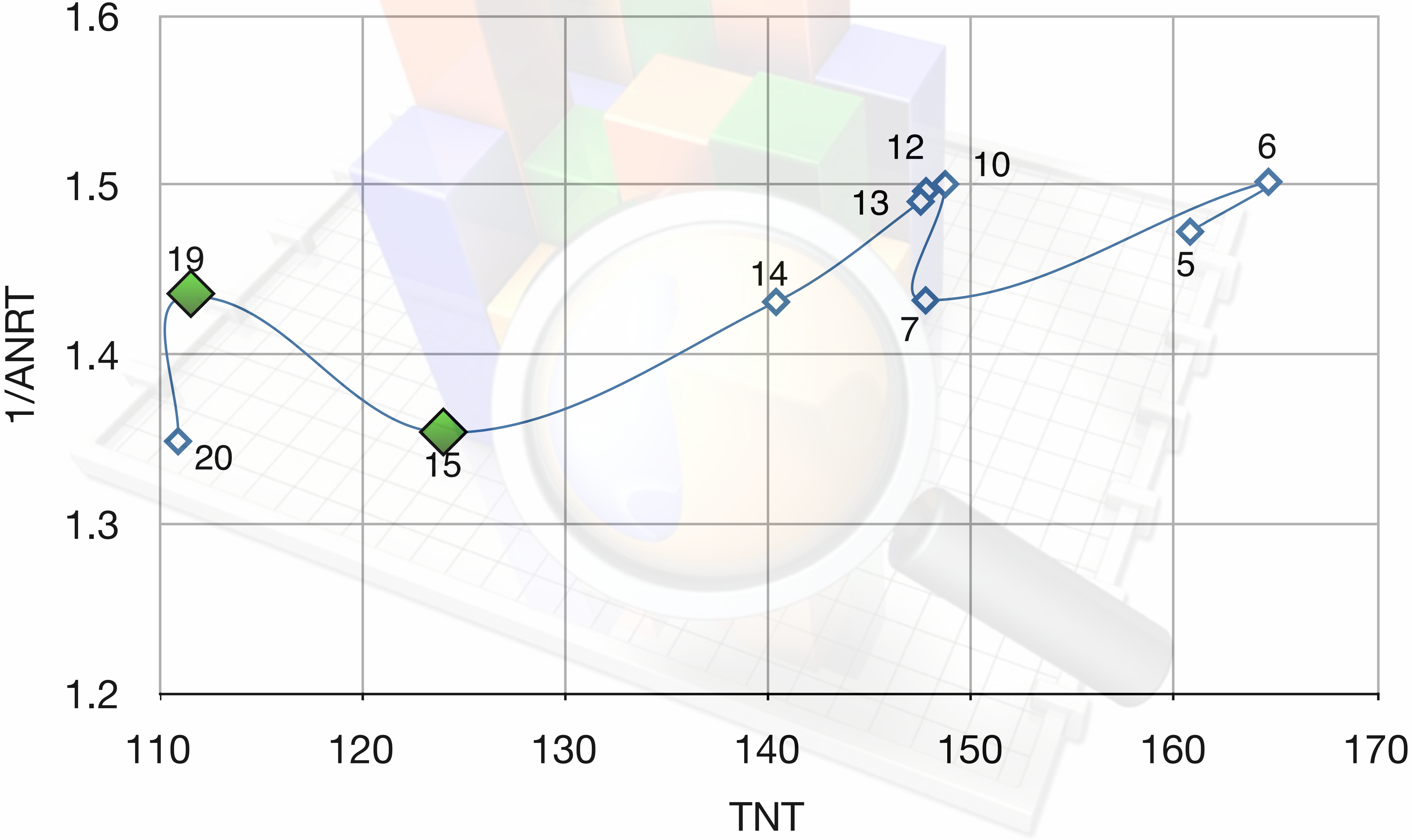
24-core systems



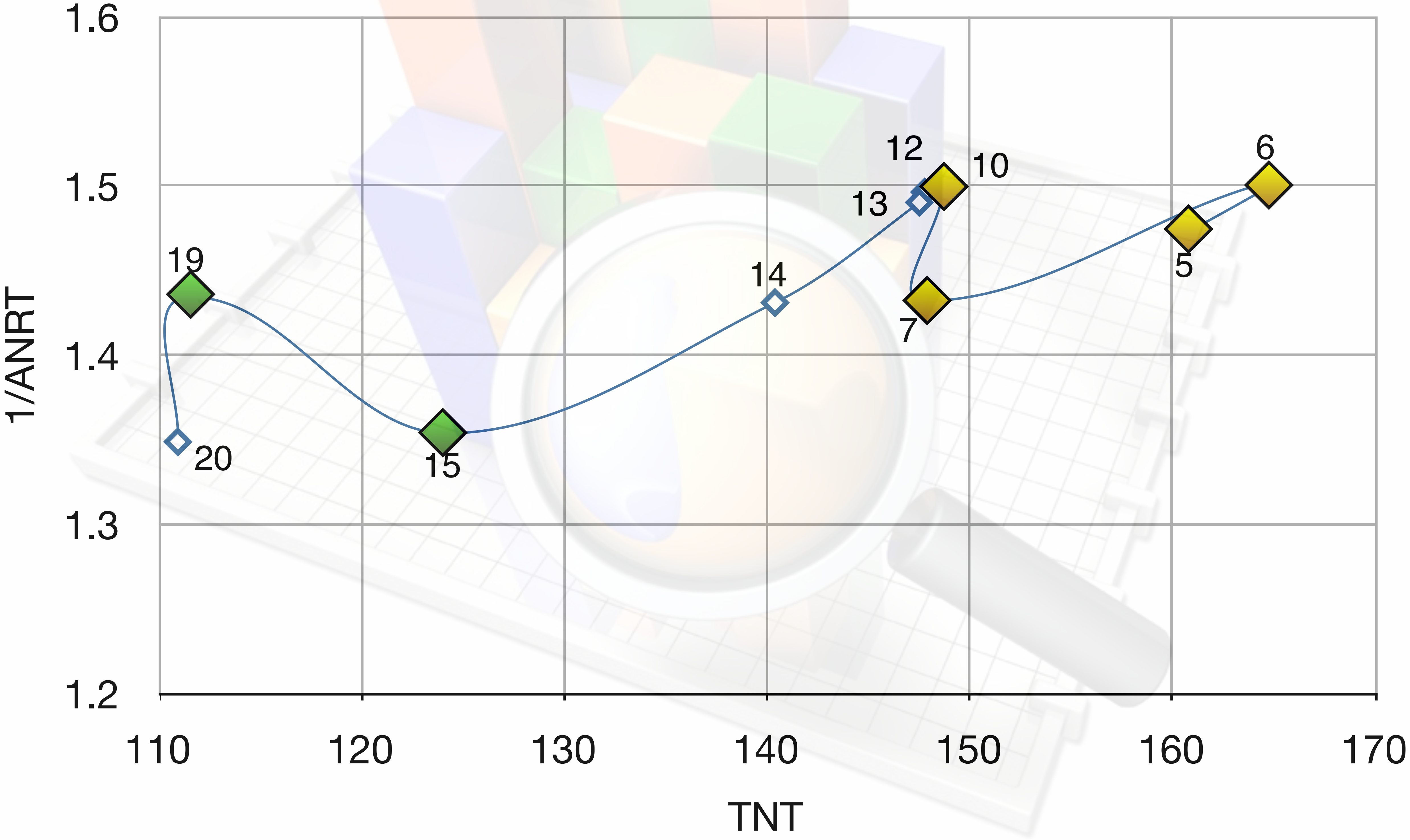
32-core systems



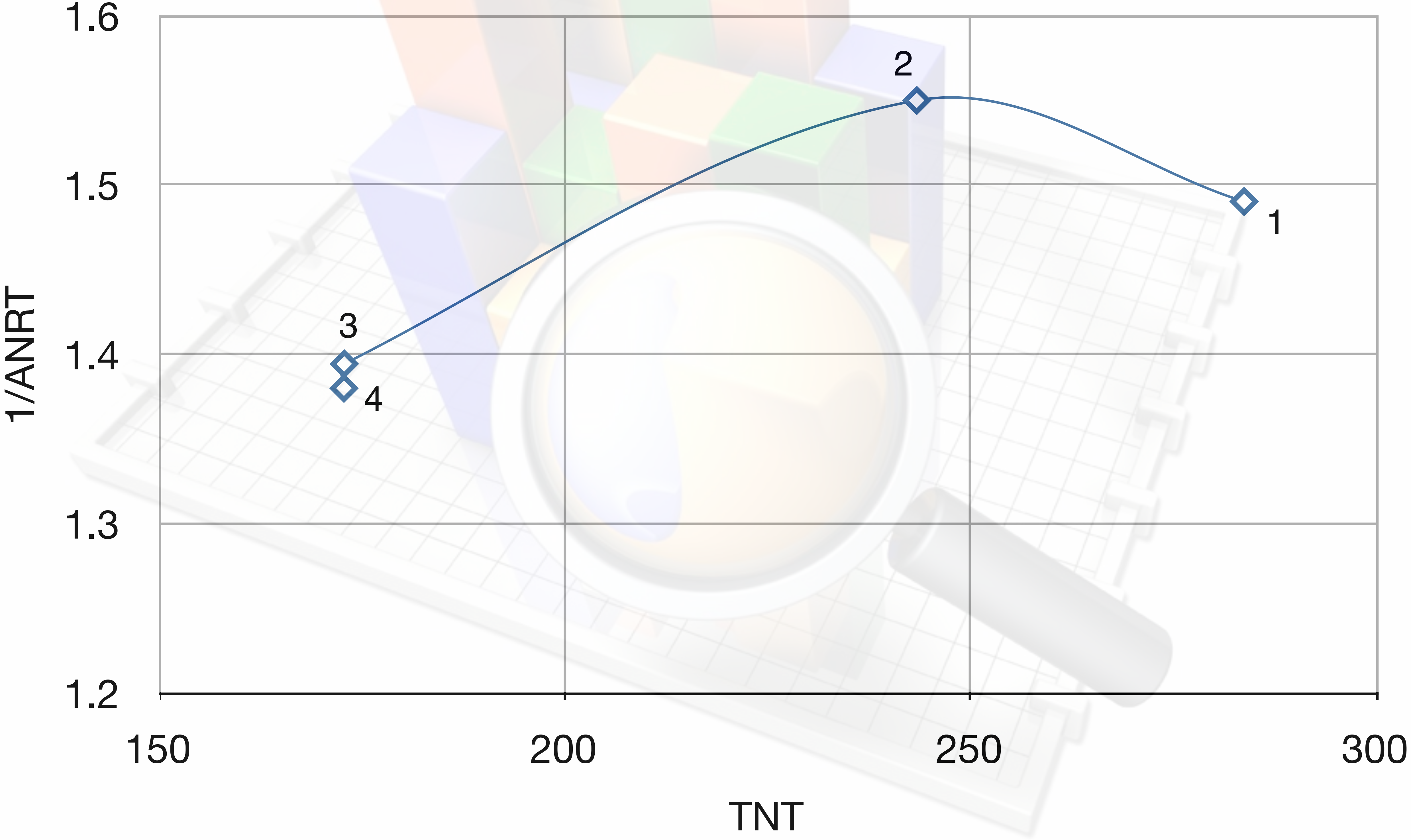
32-core systems



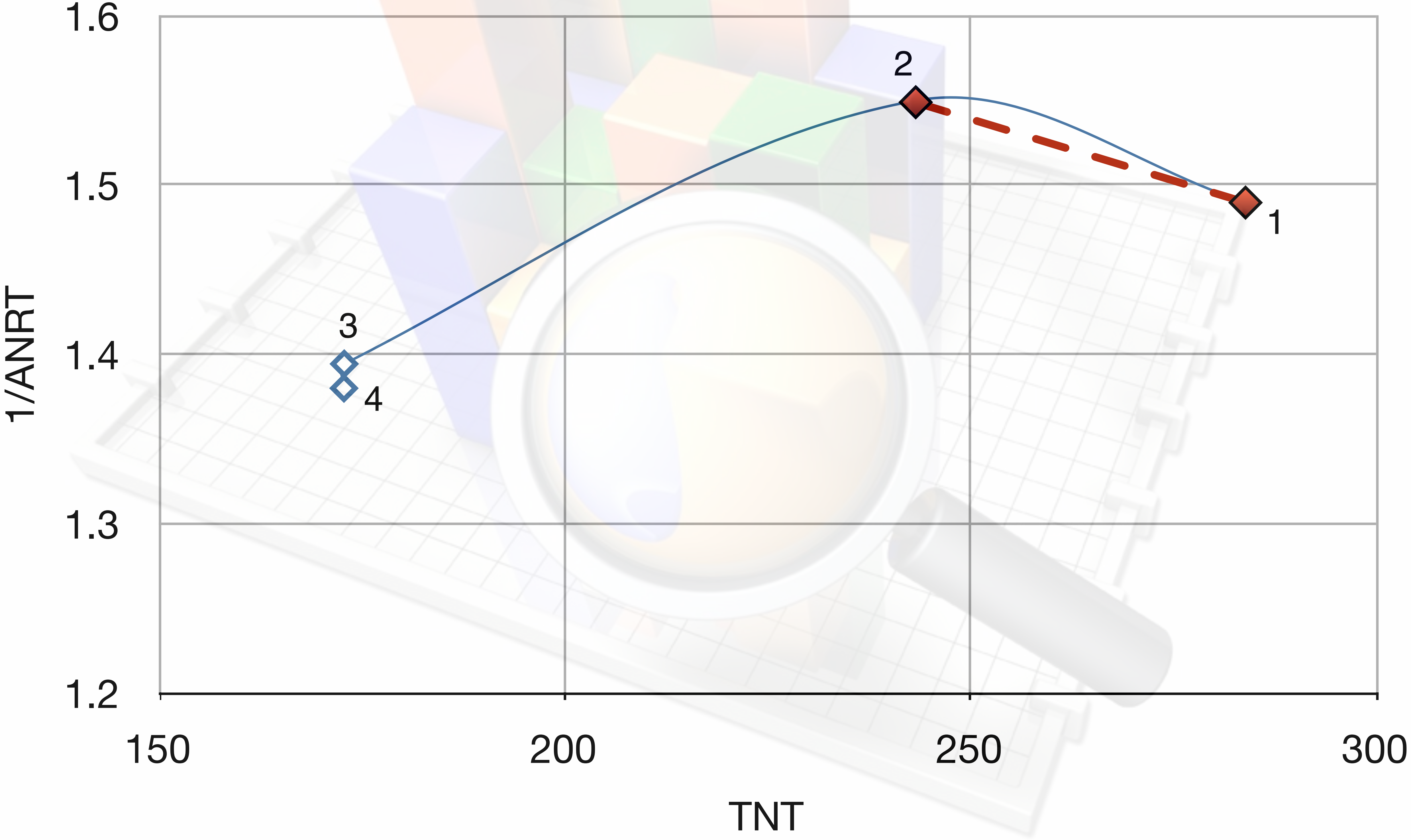
32-core systems



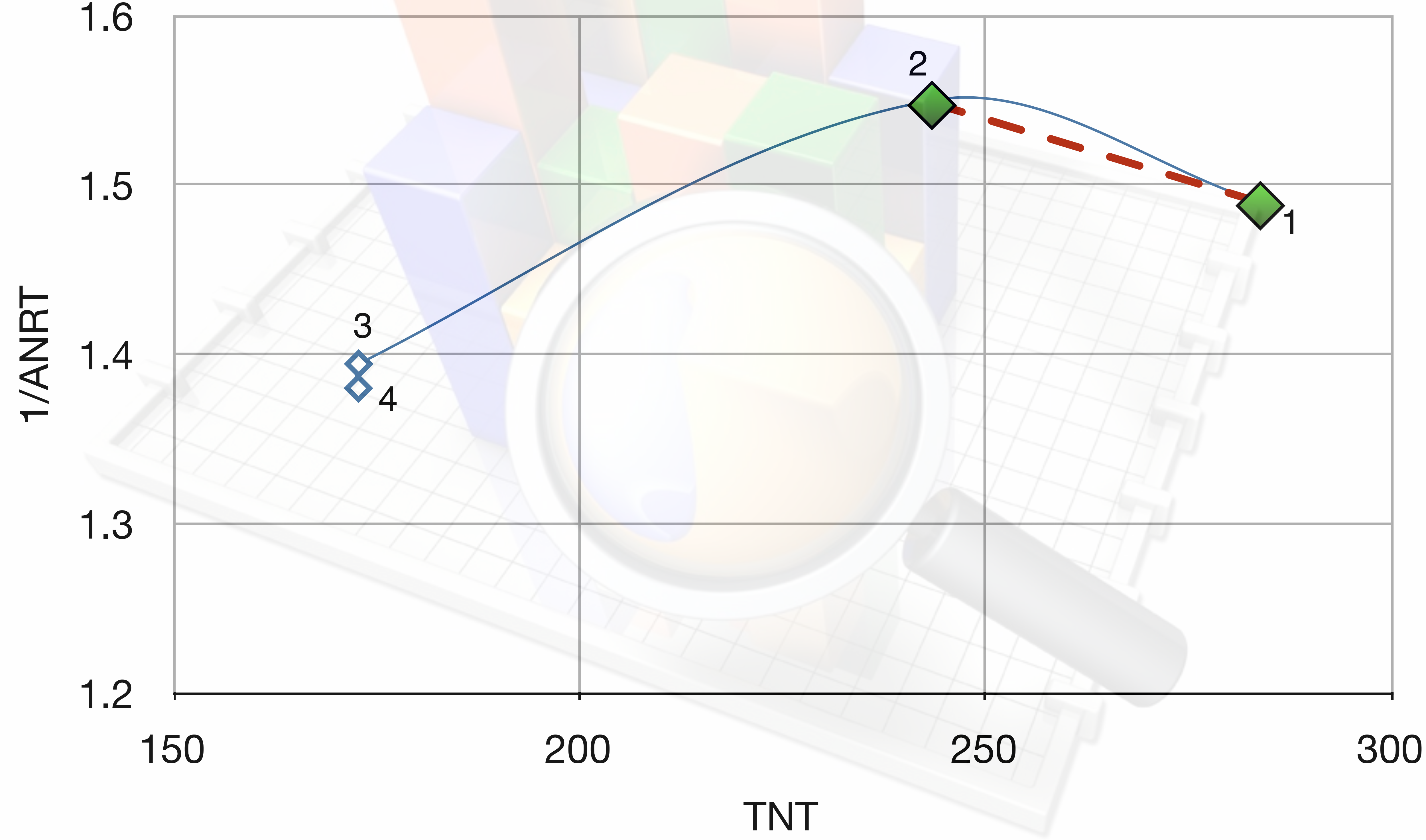
48-core systems



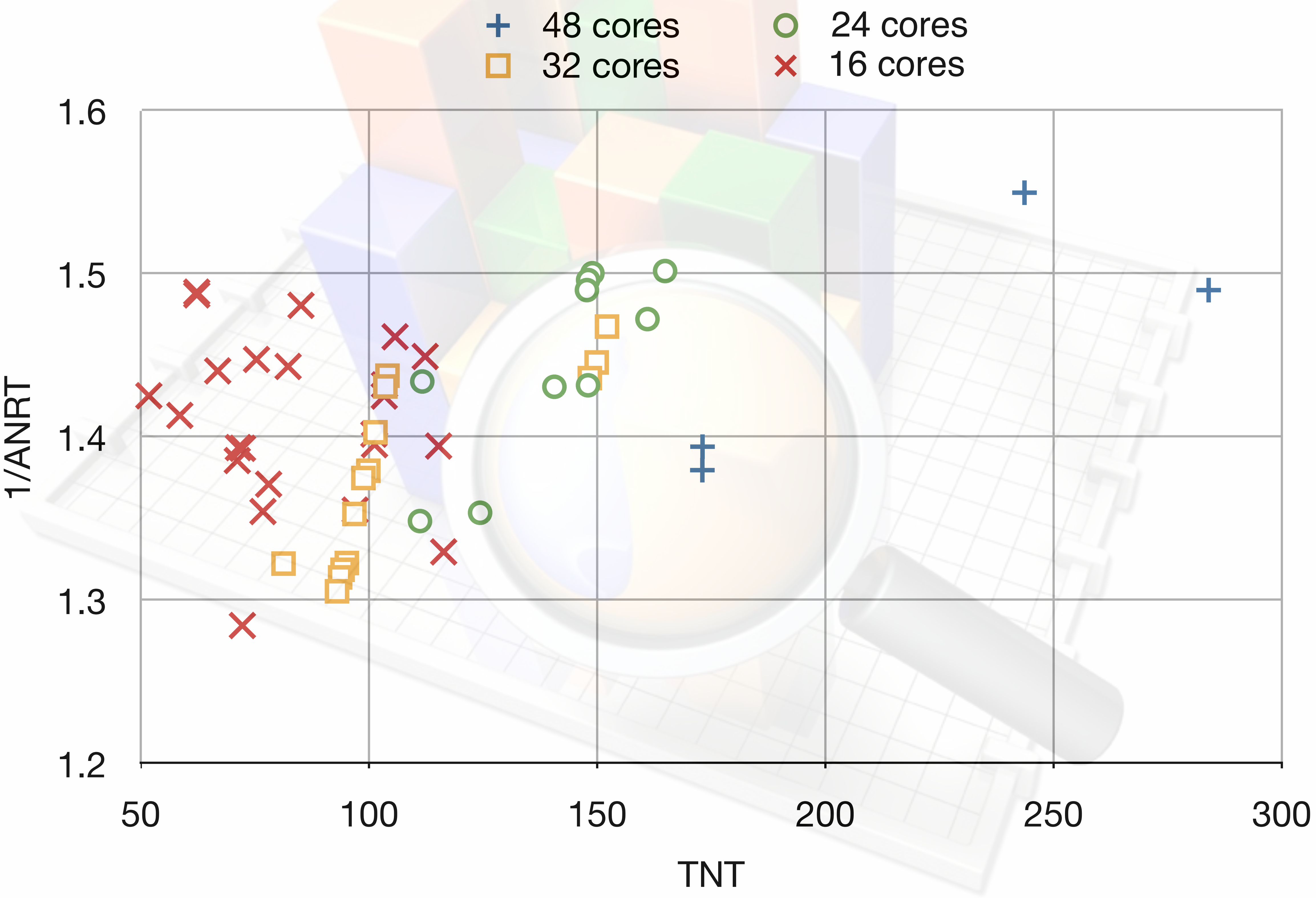
48-core systems



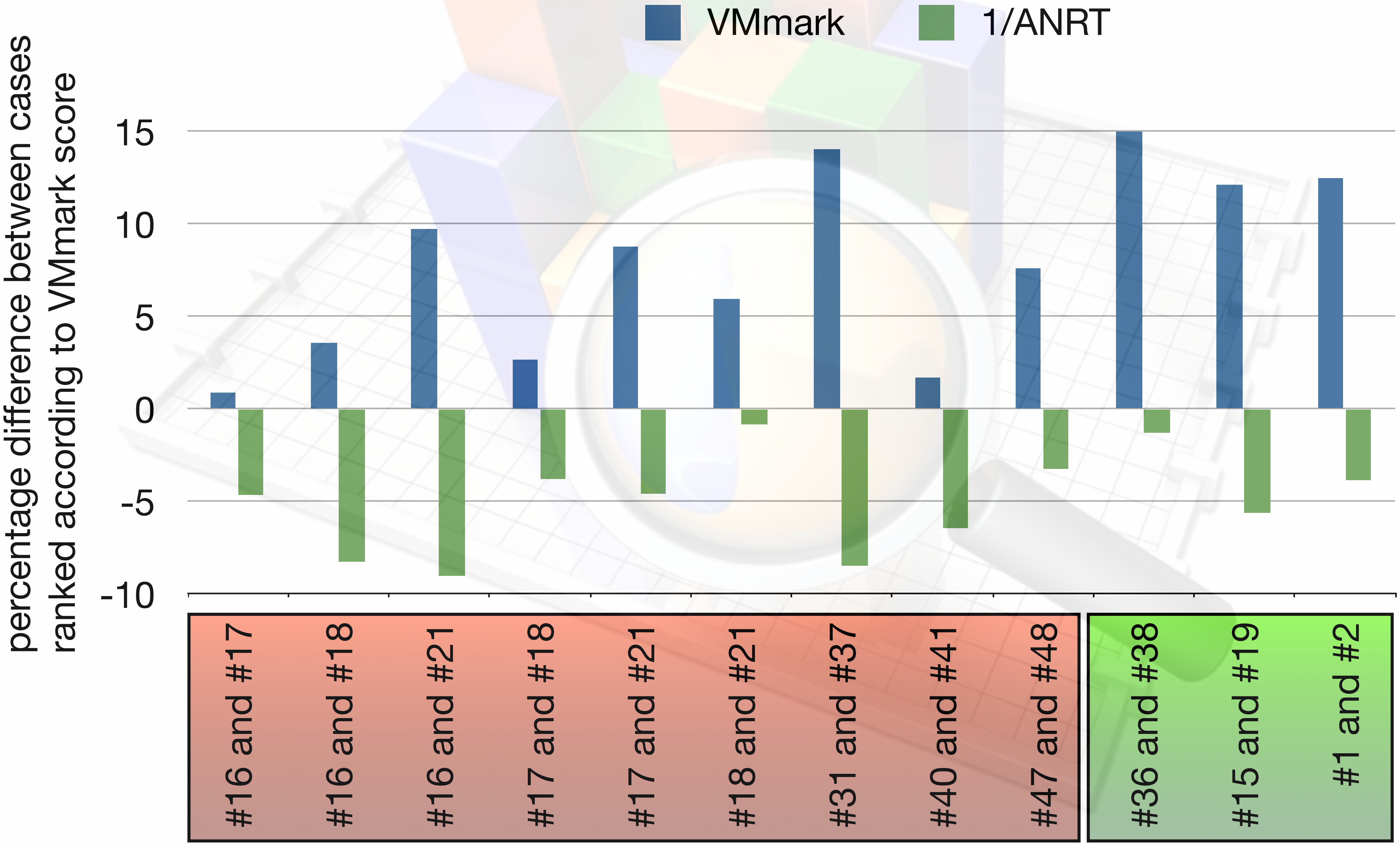
48-core systems



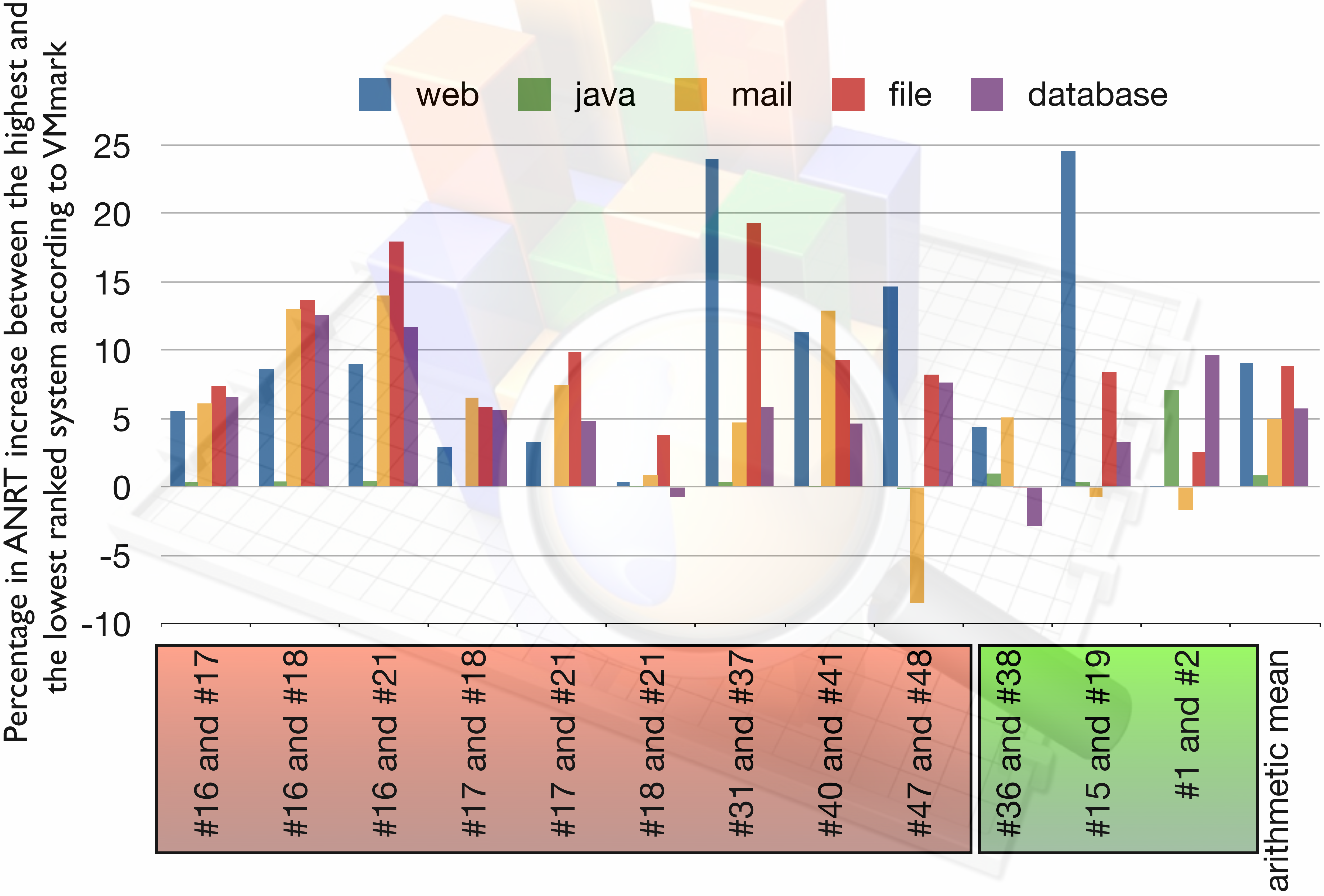
All systems


+ 48 cores
○ 24 cores
□ 32 cores
× 16 cores
1.6
1.5
1.4
1.3
1.2
1/ANRT
50
100
150
200
250
300
TNT

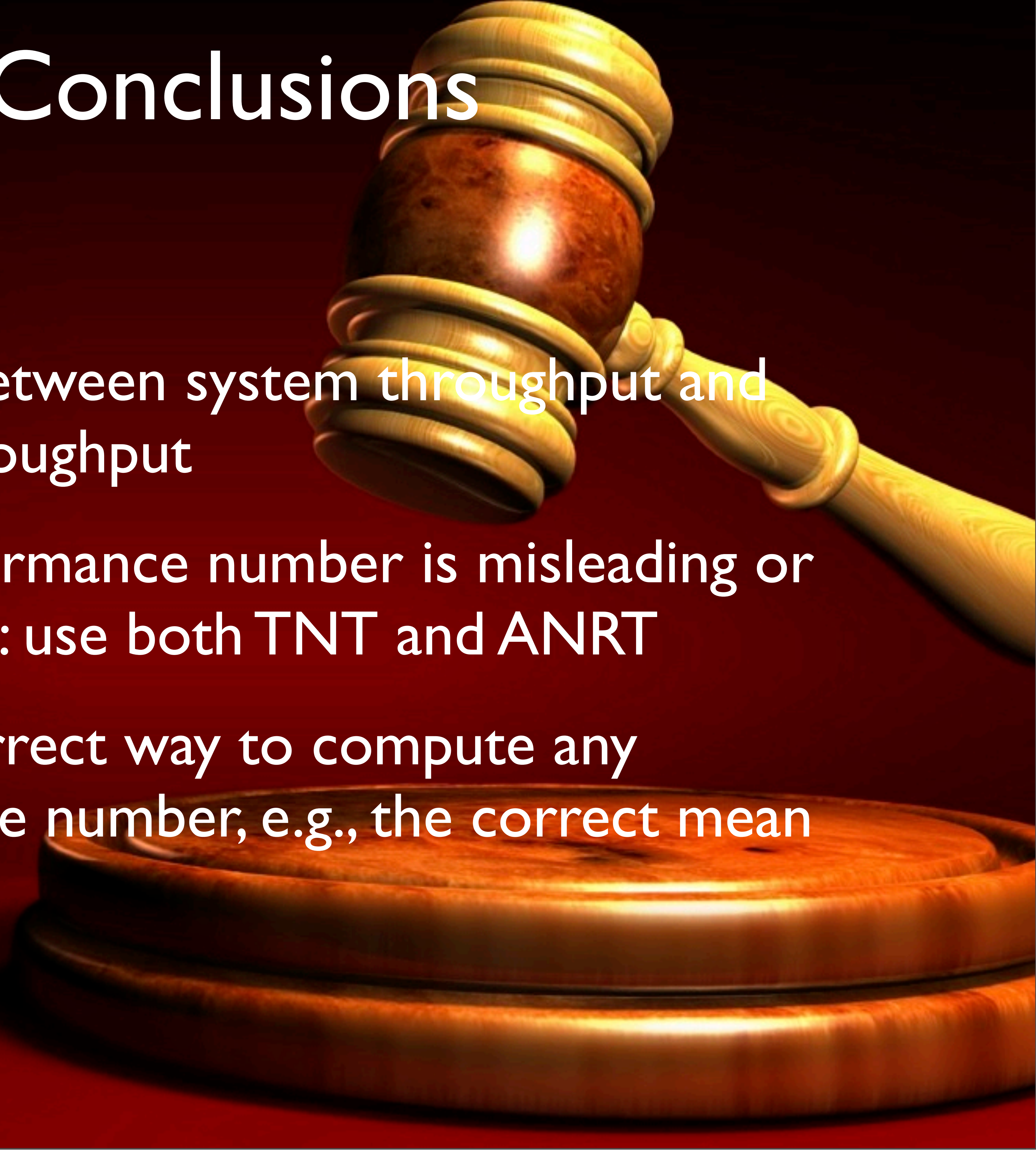
VMmark vs. ANRT



VMmark vs. ANRT



Conclusions



- Trade-off between system throughput and per-VM throughput
- Single performance number is misleading or even wrong: use both TNT and ANRT
- Use the correct way to compute any performance number, e.g., the correct mean