Software Performance & Scalability: A Cheatsheet

by Mike Kramlich

revised 2017 June 17

1. Do Nothing (why? mathwise perfect: min possible latency, cpu, mem & infinitely scalable)
2. Do Very Little (why? broadstrokes is the next most perfect & efficient thing after Do Nothing)
3. Static Not Dynamic
4. Cached
5. Distributed
6. Parallelized
7. Asynchronous
8. Incremental
9. Step Minimization
10. Paginated Results
11. Complexity (in Time or Space) Cost Optimal Algorithms (eg. O(1) over O(n) over O(n^2))
12. Event-Driven Not Polled
13. Non-Blocking IO
14. Web Page Component Request Minimization
15. Network Locality (eg. CDN's)
16. Machine Task-to-Data Locality (eg. Hadoop)
17. Precompute Predicted Requests
18. Eager Init vs Lazy Init
19. Higher CPU Clock Speed
20. Higher CPU Core Count
21. CPU Instructions Which Do More Work Per Cycle/Tock (eg. SIMD)
22. Higher Communication Bus Speed, Throughput
23. Do Tasks in Hardware Rather than Software
24. Leaner Languages & Runtimes
25. More Memory
26. Faster Memory
27. More Disk/Storage
28. Faster Disk/Storage (Seek, Read, Write)
29. Disk/Storage Defragmentation
30. Local Disk/Storage Rather Than Network Mounted
31. Higher Network Bandwidth
32. Compression of Large Transfer Payloads
33. Persistent Connections
34. Connection Pools
35. UDP not TCP
36. Minimize Chattiness of Comm Protocols
37. Pass Smaller Messages
38. Make Use of Otherwise Unused Local GPU For General/Parallel Compute Tasks
39. Make Use of Cloud Computing Services (eg. AWS)
40. Tuning OS Parameters
41. Custom OS Kernel Builds
42. Non-Virtual OS Instance
43. RTOS (if require a guarantee & hard upper bound on latency of action in response to event)
44. No OS (yes extreme but consider case of microcontroller where only 1 master program req)
45. Pass and Store Diffs Rather Than Complete Snapshots
46. Client-Server Architecture (eg. benefit: long startup init tasks done in server not clients)
47. Push Work To Client-Side Rather Than Server-Side (eg. rendering, initial input validation)

48. Local Function Calls Rather Than RPC or Web Service Requests
49. Function Memoization
50. Function Inlining
51. Loop Unrolling
52. Less Unnecessary Call Descent Depth (eg. Java/Enterprise Design Pattern Astronaut Arch)
53. Less Memory Churn and Background GC Inside Your Process
54. Object Pooling
55. Clusters
56. Queues with Worker Process/Thread Pools
57. Database Sharding
58. Database Indexing
59. Database Prepared Statements
60. Old Data Warehousing/Archiving
61. Old Metric/Event Rollups
62. Log Archiving
63. Timeout Guards
64. Buffer Size Tuning
65. Queue Size Tuning
66. Timeout Duration Tuning
67. Network Packet Size Tuning
68. Disk Page Size Tuning
69. Memory Page Size Tuning
70. Cache Size Tuning
71. Cache Eviction/Expiration Policy Tuning
72. Avoid Need To Mitigate Risk Of Hardware Failures Due To Vibration/Shock
73. Avoid Need To Mitigate Impact Of Environmental Radiation (eg. cosmic x-rays)
74. Reduce Physical Volume (eg. consider design impact on smartphones and data centers)
75. Reduce Physical Mass (eg. consider compute capabilities & cost impact on space payloads)
76. Reduce Power Consumption (eg. smartphones, laptop battery life, data centers)
77. Reduce Heat Emission (eg. impacts cooling reqs thus total cost/complexity of data centers)
78. Keep Hardware Cool, Especially Processors
79. Commodity Priced Hardware Rather Than Vendor Monopoly/Patented/Unique Hardware
80. Later/Recent Generation Hardware Models (in general: more optimized than earlier/older)
81. Decrease Max/Mean/Min Time Before Detection of Failure
82. Upstream DoS Throttling/Filtering
83. User Request Throttling
84. Automatic Load Balancing (esp smarter)
85. Off-Peak Scheduling of Tasks When Possible
86. Encourage/Require Users To Spread Access/Requests/Workloads More Evenly Over Time
87. No Encryption
88. Minimize Core/Thread Context Switching
89. Avoid Mem/Disk Paging/Swapping, Especially Thrashing
90. Higher OS Scheduling Priority For Processes Directly Responding To Live User Commands
91. Avoid Lock Contention
92. Textual UI's and CLI's Rather Than GUI's
93. Text Rather Than Images and Static Images Rather Than Video, Audio or Animations
94. Vector Illus/Anims/UI's Rather Than Bitmaps To Minimize Disk/Net/Mem Footprint
95. JSON/CSV/etc Rather Than XML (whose impls often cause higher latency or mem use)
96. Prefer Precise, Reactive, Resource Sipping Tools (eg. latency/mem/cpu of vim over Eclipse)
97. Automated Rather Than Manual (eg. no staff approval queue for content, just filter and flag)
98. Make Light Speed in Vacuum Your Only Remaining Latency Bottleneck Where Possible

This document is a personal cheatsheet list of rules-of-thumb, impact factors, ideas, patterns, strategies and techniques which can be used to improve, or at least to sculpt the performance or scalability of computing systems. Whether measured overall and universally applicable, or, merely temporarily apparent from the standpoint of any particular user, viewer or stakeholder, or as measured by one particular goal metric — possibly at the expense of worsening others.

This is not a reference manual or textbook. The traditional purpose of a cheatsheet is to remind a reader of topics they have already studied and thus should know. And to cram as much information as possible onto the smallest amount of paper.

Some of the items in this list contribute also to system availability, correctness, cost (esp lifetime TCO) or the intangible quality of the user experience. Often the effects overlap and the lines blur. As with any element in software engineering there are trade-offs to consider. For example, sometimes its better to init eagerly and sometimes init lazily. Prediction-based optimizations can help or hurt: they can help at a coarse-grained level overall yet also hurt in fine-grained individual cases. Caching can shrink latency but risks showing stale or inconsistent results. And it always helps to prioritize based on actual bottlenecks. Takes judgment to know why or how to apply any of these.

*Note that some items may not at first appear to impact performance or scalability, at least not directly, but if you zoom out a bit in your mind and consider their downstream impacts or their impacts on the overall solution, they do. For example, radiation and vibration don't necessarily *directly* hurt latency or workload capacity, but *can* cause hardware failure or data corruption, which in turn will hurt your system's correctness, availability, latency, throughput, and service lifetime. Even having features present in a design in order to *mitigate* these risks can then in turn penalize performance, however small: think about synchronization, redundant writes or error correction. Or place a drag on scaling: think about increased hardware costs to serve any given level of traffic/workload or data size, or the increased cognitive burden on engineers of having to support more complex architectures, or the increased mass and volume required to add shielding or shock absorption. Think about the need to have more or better — and therefore also more rare and expensive — engineers to design/build/support such a system. Think about bang per buck. Everything is connected.*

Also note many of the items overlap partially with one another, or can be said to be more specialized cases of other items in the list (eg. function memoization is a special case of caching.) Even in those cases there is a distinction noteworthy enough to warrant separate treatment here. I've tried to position related items near each other.

This list is a work in progress and I'll be revising it over time. Because it is part of my notes on the craft. As I remember or hear of more techniques I'll add them and publish a new version of this doc. I am considering whether to expand it by adding more detail such as concrete examples, before-vs-after comparisons with metrics, and maybe one day fleshing this all out into a book. If you'd like to see that, and especially if you'd be willing to pay for it, let me know.

Feel free to suggest additions or changes!

my email: groglogic@gmail.com

my resume:
https://synisma.neocities.org/resume_Mike_Kramlich__Software_Engineer.pdf

Feedback thanks & suggestions:
Alan Robertson (Assimilation), Brian Pontarelli (Inversoft), Taylor Deehan (MLB)

TODO decide whether to add these ideas and how to label them:

- do your compute in otherwise idle cpu moments of very large numbers of volunteer/infected hosts

- code/compile such that while running the process reads from core/cpu resident cache as much as possible

- config a (critical & cpu-bound) process to run with max priority (eg. min niceness) by the OS scheduler