VERIFYING THE INTEGRITY AND RELIABILITY OF DATA-INTENSIVE WEB-SERVICES.

A Case Study
波特兰，

REAL WORLD USAGES
In Welsh: “I am not in the office at the moment. Send any work to be translated.”
THE MSR MACHINE TRANSLATION (MT) SYSTEM

User makes a request to translate a web-page

Received by MT web-service

Sentence and word breaking

Sentence Translation

Translation engine and models

Document re-creation

Back to user.
COMMON APPROACH TO TESTING WEB-SERVICES

- Verify individual components first.
- Verify components in integration.
- Stress and performance testing.
- Mean-time-to-failure.
- Identify maximum sustainable traffic.
- Capacity planning.
DATA-INTENSIVE WEB-SERVICES

- Obvious performance impact.
- A different set of failure points.
- Bottlenecks.
- Data integrity and availability.
- Growth patterns.
- Upgrade/versioning.
THE DATA-INTENSIVE COMPONENTS

- Translation models hosted by a sub-set of the MT web-services.
- Data size: multiple GBs.
- Loaded at run-time, and memory-mapped files.
- Updated for every scrum cycle.
MEMORY-MAPPED FILES

- A useful solution to increase performance.
- Works well for most scenarios.
- More not always better: increased contention/error rate for large data sizes.
- Run tests to measure optimal size.

<table>
<thead>
<tr>
<th></th>
<th>With 2 instances of memory mapped file</th>
<th>With 4 instances of memory mapped file</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput, in words per second</td>
<td>5,000</td>
<td>8,000</td>
</tr>
<tr>
<td>Percentage of requests failing</td>
<td>1.6%</td>
<td>13.5%</td>
</tr>
</tbody>
</table>
MEMORY LEAKS
SIDE-EFFECT OF HANDLE COUNT INCREASE
OVERHEAD OF INDIVIDUAL COMPONENTS

- In a distributed system, it is important to minimise communication or processing overhead.
- Identify primary contributor to a bottleneck.
- Tests run regularly measuring throughput at different layers.

![Graph of Translation Throughput Numbers](image)
STAYING WITH IT.

- Run perf tests regularly.
- More quickly identify effect of rolling changes.
IMPACT OF GARBAGE COLLECTION

- For data-intensive applications/services, any overhead is performance hit.
- Side-effect of managed cache: time in GC.
- Useful to measure: percentage time in GC, large object heap size,
FLAVOURS OF GARBAGE COLLECTION

- Difference between Workstation and Server-GC:
CONCLUSIONS

- Squeeze throughput to the utmost at each layer.
- Stay on top of version/build changes.
- Keep It Simple.