ExPERT: Pareto-efficient task replication on grids and a cloud

Orna Agmon Ben-Yehuda$^1$ Assaf Schuster$^1$
Artyom Sharov$^1$ Mark Silberstein$^1$ Alexandru Iosup$^2$

$^1$Department of Computer Science
Technion — Israel Institute of Technology

$^2$Faculty of Engineering, Mathematics and Computer Science (EWI)
TU Delft

IPDPS, May 2012
Owner goals – minimize:
* Operational costs (energy)
* Effective load

User goals – minimize:
* Makespan
* Cost

Policy
Enforcing QoS
Costs

Resource OWNER

Workload
Paying for resources or QoS
Strategy (Declarations, Resource Usage)

Resource USER
Owner goals – minimize:
* operational costs (energy)
* effective load

User goals – minimize:
* Makespan
* Cost

An environment of uncertainty: Will the task fail on the unreliable resource? Which system to use?
In the Beginning...

Unreliable queue \(\rightarrow\) Unreliable Pool

- \#machines < \#unfinished tasks.
- \(D\) - instance deadline.
- No replication (replication is inefficient).
Using the Same Strategy After the Tail Starts

#machines > #unfinished tasks

The tail is wagging the dog...

Agmon Ben-Yehuda, Schuster, Sharov, Silberstein, Iosup
ExPERT
Replication - the User’s Bank of $NTDM_r$ Strategies

First N Tail Instances

Unreliable queue

Unreliable Pool

Failure/Timeout

Success

Instance N+1 in Tail

Reliable Queue

Reliable Pool

Success

- $D$ - instance deadline, $T$ - replication time
- Reliable machine used to ensure task completion
- $N$ tail instances at most on unreliable resources
- $M_r$ - max ratio of reliable to unreliable resources
Replication - the User’s Bank of $NTDM_r$ Strategies

- First N Tail Instances
  - Unreliable queue
  - Instance N+1 in Tail
    - Reliable Queue

- Unreliable Pool
  - Failure/Success
  - T
  - D
  - Timeout

- Reliable Pool
  - Success

- $D$ - instance deadline, $T$ - replication time
- Reliable machine used to ensure task completion
- $N$ tail instances at most on unreliable resources
- $M_r$ - max ratio of reliable to unreliable resources
Example: Number of unreliable instances $N = 3$

- Replication wastes work!
The user cares about multi-objective optimization:

- $\langle Cost \rangle$ - Mean $\frac{\text{cost}}{\text{task}}$ or $\frac{\text{tail-cost}}{\text{tail-task}}$
- $\langle MS \rangle$ - Mean makespan or tail makespan.

Each user may have her own objective, pending on those values:

- Below minimal makespan: $\langle MS \rangle < Const$
- As fast as possible: $\min \langle MS \rangle$
- Below max budget: $\langle Cost \rangle < Const$
- As cheap as possible: $\min \langle Cost \rangle$
- Best price for the goods: $\min \langle Cost \rangle \langle MS \rangle$
- Any other function of means: $\langle Cost \rangle, \langle MS \rangle \ldots$
Users who do not optimize well behave irrationally and are hard to predict.
Rational users can optimize general utility function.
Towards the final goal of manipulating users to save energy
Get **user additional data** (costs, reliable pool times).
Get unreliable resource statistics (trace analysis).
Compute a Pareto frontier for $\langle \text{Cost} \rangle, \langle \text{MS} \rangle$. 

ExPERT 15/33
Solution Concept - Step 3

Estimate $\langle Cost \rangle, \langle MS \rangle$ for each strategy in the search space:

- For every working point, the *ExPERT Estimator* computes several random realizations on the basis of the statistic characterization.

- The average makespan and cost over these realizations are used as the expectation values $\langle Cost \rangle, \langle MS \rangle$.
Estimate $\langle Cost \rangle, \langle MS \rangle$ for each strategy in the search space.

- Filter out dominated strategies.
- Keep frontier composed of non-dominated strategies.
Estimate $\langle Cost \rangle, \langle MS \rangle$ for each strategy in the search space.
Filter out dominated strategies.
Keep frontier composed of non-dominated strategies.
- Estimate $\langle Cost \rangle, \langle MS \rangle$ for each strategy in the search space.
- Filter out dominated strategies.
- Keep frontier composed of non-dominated strategies.
Estimate \( \langle Cost \rangle, \langle MS \rangle \) for each strategy in the search space.

Filter out dominated strategies.

Keep frontier composed of non-dominated strategies.
Estimate \(\langle Cost \rangle, \langle MS \rangle\) for each strategy in the search space.

Filter out dominated strategies.

Keep frontier composed of non-dominated strategies.
Choose **optimal** strategy according to user utility (by expectation value).

Get $N, T, D, M_r$ params for the desired strategy.
Apply strategy: Feed \( N, T, D, M_r \) params as input to the user scheduler and deploy tasks on the resource pools.
GridBoT:

- Supplies a unified front-end to multiple grids and clouds using their local resource management infrastructure.
- Employs dynamic run-time scheduling and replication strategies to execute BoTs in multiple environments simultaneously.

A BoT trace holds a line per task with the following fields:

- Status (failed/succeeded)
- Runtime: only for successful tasks.
- Wait time: from submitting to starting running. May be unavailable for failed tasks.

Result time = Runtime + Wait time

UW-M: Condor cluster of University of Wisconsin-Madison.
Characterize the Unreliable Resource

- \#ur: the effective size of the unreliable pool.
- \( F(t, t') = \text{reliability}(t') \cdot F_s(t) \), CDF of result turnaround time, on the basis of:
  - \( F_s(t) \): the measured CDF of result turnaround time \((t)\) of successful tasks.
  - \( \text{reliability}(t') \): the fraction of successful tasks as a function of the time since the BoT started \( t' \).
Local optimization is not trivial.

Unoptimized strategies are wasteful.
Optimizing a General User Utility Function along the Pareto Frontier

Utility function choice can be postponed till the user is aware of the cost-makespan tradeoff. It does not have to be expressed.
Validation and Evaluation

ExPERT Estimator
Validation and Evaluation

ExPERT Estimator

Simulator Wrapper
Validation and Evaluation

ExPERT
Estimator
Simulator
Wrapper

Validation

Experiments on real grids and a cloud (EC2)
Validation and Evaluation

Simulator Wrapper

Validated Simulator

ExPERt Estimator

Validation

Experiments on real grids and a cloud (EC2)
Validation and Evaluation

- Validated Estimator
- Validated Simulator
- ExPert Simulator Wrapper
- Validation
- Experiments on real grids and a cloud (EC2)
Validation and Evaluation

ExPERt Estimator

Validation

Experiments on real grids and a cloud (EC2)

Simulator Wrapper

Validated Simulator

Validated Estimator

Performance Compared with Static Strategies Scheduling
Validation and Evaluation

ExPERT Simulator Wrapper

Validated Estimator

Validated Simulator

Performance Compared with Static Strategies Scheduling

Insights Regarding Efficient Resource Use

Validation

Experiments on real grids and a cloud (EC2)

Evaluation

Agmon Ben-Yehuda, Schuster, Sharov, Silberstein, Iosup
Table: Resource Pools

<table>
<thead>
<tr>
<th>Reliable Pool</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technion</td>
<td>20 self-owned CPUs in the Technion.</td>
</tr>
<tr>
<td>EC2</td>
<td>20 large EC2 cloud instances.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unreliable Pool</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>UW-M</td>
<td>UW-Madison Condor pool (preempts).</td>
</tr>
<tr>
<td>OSG</td>
<td>Open Science Grid (no preemption).</td>
</tr>
<tr>
<td>UW-M + OSG</td>
<td>Combined: half ♯ur from each pool.</td>
</tr>
</tbody>
</table>
Validation: Prediction Deviation

- Relative tail makespan deviation
- Relative tail cost deviation
- Offline deviation
- Online deviation
- Mean absolute offline deviation
- Mean absolute online deviation

X: 20.38, Y: 13
X: 10.38, Y: 7.077

Agmon Ben-Yehuda, Schuster, Sharov, Silberstein, Iosup - ExPERT
Augmenting the Estimator with Static Strategies

1. AR: All to Reliable
2. TRR: all Tail Replicated to Reliable (N=0, T=0)
3. TR: all Tail to Reliable (N = 0, T = D)
4. AUR: All to UnReliable, no replication
5. B: Budget of 1$ for a BoT of 150 tasks (on average, \(\frac{2}{3} \text{ cent BoT task}\))
6. CN\(\infty\): Combine resources, no replication
7. CT0N1: Combine resources, replicate at tail with N = 1, T = 0
ExPert’s Pareto frontier dominates most static strategies.
Performance: makespan $\times$ cost, 3 $M_{r}^{\text{max}}$ values

Smaller is better. Cost $\times$ makespan ExPERT’s recommended strategy is 25% lower than second-best and at least 72% lower than third-best.

Agmon Ben-Yehuda, Schuster, Sharov, Silberstein, Iosup  ExPERT  29/33
On the Pareto frontier, usually all reliable resources are used at some point, and a significant queue is built for them.

Agmon Ben-Yehuda, Schuster, Sharov, Silberstein, Iosup  ExPERT 30/33
The free parameter $M_r$ enables efficient strategies with lower costs for the same makespan. It makes tasks wait in a queue, where they may be canceled.

Agmon Ben-Yehuda, Schuster, Sharov, Silberstein, Iosup
The $NTM_r$ strategy space is vast enough to provide user preference flexibility.

ExPERT-recommended strategies finish in two-thirds of the time and cost a quarter of commonly used static strategies.

Using ExPERT means you do not waste time or money, and you optimize your own utility function.
Questions?

Contact us at:

- Orna Agmon Ben-Yehuda ladypine@cs.technion.ac.il
- Assaf Schuster assaf@cs.technion.ac.il
- Artyom Sharov sharov@cs.technion.ac.il
- Mark Silberstein marks@cs.technion.ac.il
- Alexandru Iosup A.Iosup@tudelft.nl

Thank You!

夫人杉木