Static Worksharing Strategies for Heterogeneous Computers with Unrecoverable Failures

Anne Benoit, Yves Robert,
Arnold Rosenberg and Frédéric Vivien

École Normale Supérieure de Lyon, France
Anne.Benoit@ens-lyon.fr
http://graal.ens-lyon.fr/~abenoit

HeteroPar'2009, August 25

Problem

- Large divisible computational workload
- Single-round distribution, one-port model
- Assemblage of p different-speed computers
- Unrecoverable interruptions
- A-priori knowledge of risk (failure probability)

Goal: maximize expected amount of work done

Related work

- Landmark paper by Bhatt, Chung, Leighton & Rosenberg on cycle stealing
- Hardware failures

☺ **Fault tolerant computing (hence scheduling) becomes unavoidable**

☹ **Well, same story told since very long!**

Related work

- Landmark paper by Bhatt, Chung, Leighton & Rosenberg on cycle stealing
- Hardware failures

☺ **Fault tolerant computing (hence scheduling) becomes unavoidable**

☹ **Well, same story told since very long!**

Related work

- Landmark paper by Bhatt, Chung, Leighton & Rosenberg on cycle stealing
- Hardware failures

☺ **Fault tolerant computing (hence scheduling) becomes unavoidable**

☹ **Well, same story told since very long!**

Cycle-stealing scenario

- Big job of size W to execute during week-end
- Enroll p computers P_1 to P_p
- Assign load fraction to each P_i
- **How to compute these load fractions?**
- **How to order communications?**

- Risk increases with time
- Machines reclaimed at 8am on Monday with probability 1

Cycle-stealing scenario

- Big job of size W to execute during week-end
- Enroll p computers P_1 to P_p
- Assign load fraction to each P_i
- **How to compute these load fractions?**
- **How to order communications?**

- Risk increases **linearly** with time
- Machines reclaimed at 8am on Monday with probability 1

Cycle-stealing scenario

- Big job of size W to execute during week-end
- Enroll p computers P_1 to P_p
- Assign load fraction to each P_i
- **How to compute these load fractions?**
- **How to order communications?**

- Risk increases **linearly** with time
- Machines reclaimed at 8am on Monday with probability 1

Outline

Outline

Interruption model

$$dPr = \begin{cases} \kappa dt & \text{for} \quad t \in [0, 1/\kappa] \\ 0 & \text{otherwise} \end{cases}$$

$$Pr(w) = \min\left\{1, \int_0^w \kappa dt\right\} = \min\{1, \kappa w\}$$

Goal: maximize expected work production

Interruption model

$$dPr = \begin{cases} \kappa dt & \text{for} \quad t \in [0, 1/\kappa] \\ 0 & \text{otherwise} \end{cases}$$

$$Pr(w) = \min\left\{1, \int_0^w \kappa dt\right\} = \min\{1, \kappa w\}$$

Goal: maximize expected work production

Rules of the game

- Single-round, no overlap, one-port communications
- Homogeneous network
- Different-speed computers

- Failure-rate per unit-load **communication**

$$z = \frac{\kappa}{bw}$$

- Failure-rate per unit-load **computation** by computer P_i

$$x_i = \frac{\kappa}{speed_i}$$

Rules of the game

- Single-round, no overlap, one-port communications
- Homogeneous network
- Different-speed computers

- Failure-rate per unit-load **communication**

$$z = \frac{\kappa}{\text{bw}}$$

- Failure-rate per unit-load **computation** by computer P_i

$$x_i = \frac{\kappa}{\text{speed}_i}$$

With two computers (1/2)

P_1 $\underline{z\ Y}$ $\underline{x_1\ Y}$

- First send P_1 a chunk of size Y:
  $E_1 = Y(1 - (z + x_1)Y)$
- Then send P_2 the remaining load (of size $W - Y$):
  $E_2 = (W - Y)(1 - (zW + x_2(W - Y))$
- Total expectation:
  $E(Y) = E_1 + E_2$

With two computers (1/2)

P_1 $\quad$ $z\,Y$ $\quad$ $x_1\;Y$

P_2 $\quad$ $z\,(W-Y)$ $\quad$ $x_2\,(W-Y)$

- First send P_1 a chunk of size Y:
  $E_1 = Y\,(1-(z+x_1)\,Y)$
- Then send P_2 the remaining load (of size $W-Y$):
  $E_2 = (W-Y)\,(1-(zW+x_2(W-Y))$
- Total expectation:
  $E(Y) = E_1 + E_2$

With two computers (1/2)

P_1 $\quad z\,Y \quad x_1\,Y$

P_2 $\quad z\,(W-Y) \quad x_2\,(W-Y)$

- First send P_1 a chunk of size Y:
  $E_1 = Y\,(1-(z+x_1)\,Y)$
- Then send P_2 the remaining load (of size $W-Y$):
  $E_2 = (W-Y)\,(1-(zW+x_2(W-Y))$
- Total expectation:
  $E(Y) = E_1 + E_2$

With two computers (2/2)

$$E(Y) = Y\left(1 - (z + x_1)Y\right) + (W - Y)\left(1 - (zW + x_2(W - Y)\right)$$

$$E(Y) = W - (z + x_2)W^2 - (z + x_1 + x_2)Y^2 + (z + 2x_2)WY$$

$$Y^{(opt)} = \frac{z + 2x_2}{2(z + x_1 + x_2)}W$$

$$E_{opt}(W, 2) = E(Y^{(opt)}) = W - \left(\frac{4x_1x_2 + 4(x_1 + x_2)z + 3z^2}{4(x_1 + x_2 + z)}\right)W^2$$

Symmetric in x_1 and x_2

$\Rightarrow$ ordering of the communications has **no impact**

With two computers (2/2)

$$E(Y) = Y\,(1-(z+x_1)Y) + (W-Y)\,(1-(zW+x_2(W-Y)))$$

$$E(Y) = W - (z+x_2)W^2 - (z+x_1+x_2)Y^2 + (z+2x_2)WY$$

$$Y^{(\text{opt})} = \frac{z+2x_2}{2(z+x_1+x_2)}W$$

$$E_{\text{opt}}(W,2) = E(Y^{(\text{opt})}) = W - \left(\frac{4x_1x_2 + 4(x_1+x_2)z + 3z^2}{4(x_1+x_2+z)}\right)W^2$$

Symmetric in x_1 and x_2

⇒ ordering of the communications has **no impact**

With two computers (2/2)

$$E(Y) = Y\left(1-(z+x_1)Y\right) + (W-Y)\left(1-(zW + x_2(W-Y)\right)$$

$$E(Y) = W - (z+x_2)W^2 - (z+x_1+x_2)Y^2 + (z+2x_2)WY$$

$$Y^{(\mathrm{opt})} = \frac{z+2x_2}{2(z+x_1+x_2)}W$$

$$E_{\mathrm{opt}}(W,2) = E(Y^{(\mathrm{opt})}) = W - \left(\frac{4x_1x_2 + 4(x_1+x_2)z + 3z^2}{4(x_1+x_2+z)}\right)W^2$$

Symmetric in x_1 and x_2

$\Rightarrow$ ordering of the communications has **no impact**

Extra rule: distribute entire load

- Total load W small enough so that we distribute it entirely
- Quite reasonable but dramatic impact on solution

Definition

DISTRIB(p): compute $E_{\text{opt}}(W, p)$, the optimal value of expected total amount of work done when distributing entire workload $W \leq \frac{1}{z+\max(x_i)}$ to the p remote computers

A sufficient condition

Proposition

If $W \leq \frac{1}{z+\max(x_i)}$, there is a non-zero probability that the last computer does not fail before or during its computation

Proof

- last computer P_i can start computing at time-step Y/bw, where $Y \leq W$ is the total load sent to all preceding computers
- introducing idle times cannot improve solution: failure risk grows with time
- then P_i needs V/speed_i time-steps to execute its own chunk of size V, where $Y + V \leq W$

Outline

Optimal solution

Theorem

When $x_i = x$ (identical speeds), the optimal solution to $\textsc{Distrib}(p)$ *is obtained with same size chunks (hence of size $\frac{W}{p}$), and*

$$E_{opt}(W,p) = W - \frac{(p+1)z + 2x}{2p} W^2$$

- Closed-form formula ☺
- Proof by induction

Proof (1/2)

- Let $f_p = \frac{(p+1)z+2x}{2p}$
- We prove by induction on p that $E_{opt}(W, p) = W - f_p W^2$, with same size chunks
- Case $p = 1$, $f_1 = z + x$, $E_{opt}(W, 1) = W(1 - (z + x)W)$, OK
- From n to $n+1$ computers:
  - chunk sent to P_{n+1} of size $W - Y$
  - by induction $E_{opt}(Y, n) = Y(1 - f_n Y)$, with chunk sizes $\frac{Y}{n}$
  - for $n+1$ computers, we have

$$E(Y) = Y(1 - f_n Y) + (W - Y)(1 - zW - x(W - Y))$$

Proof (1/2)

- Let $f_p = \frac{(p+1)z+2x}{2p}$
- We prove by induction on p that $E_{\text{opt}}(W,p) = W - f_p W^2$, with same size chunks
- Case $p = 1$, $f_1 = z + x$, $E_{\text{opt}}(W,1) = W(1-(z+x)W)$, OK
- From n to $n+1$ computers:
  - chunk sent to P_{n+1} of size $W - Y$
  - by induction $E_{\text{opt}}(Y,n) = Y(1 - f_n Y)$, with chunk sizes $\frac{Y}{n}$
  - for $n+1$ computers, we have

$$E(Y) = Y(1 - f_n Y) + (W-Y)(1 - zW - x(W-Y))$$

Proof (1/2)

- Let $\mathrm{f}_p = \frac{(p+1)\mathrm{z}+2\mathrm{x}}{2p}$
- We prove by induction on p that $E_{\mathrm{opt}}(W, p) = W - \mathrm{f}_p W^2$, with same size chunks
- Case $p = 1$, $\mathrm{f}_1 = \mathrm{z} + \mathrm{x}$, $E_{\mathrm{opt}}(W, 1) = W(1 - (\mathrm{z}+\mathrm{x})W)$, OK
- From n to $n+1$ computers:
  - chunk sent to P_{n+1} of size $W - Y$
  - by induction $E_{\mathrm{opt}}(Y, n) = Y(1 - \mathrm{f}_n Y)$, with chunk sizes $\frac{Y}{n}$

  - for $n+1$ computers, we have

$$E(Y) = Y(1 - \mathrm{f}_n Y) + (W - Y)(1 - \mathrm{z}W - \mathrm{x}(W - Y))$$

Proof (2/2)

- $E(Y) = W - (z + x)W^2 - (f_n + x)Y^2 + (z + 2x)WY$
- $Y^{(opt)} = \frac{z+2x}{2(f_n+x)} W$
- $E_{opt}(W, n+1) = E(Y^{(opt)}) = W - \alpha W^2$, where $\alpha = z + x - \frac{(z+2x)^2}{4(f_n+x)}$
- By induction, $f_n + x = \frac{(n+1)z+2x}{2n} + x = \frac{(n+1)(z+2x)}{2n}$
- Finally, $\alpha = z + x - \frac{n(z+2x)}{2(n+1)} = \frac{(n+2)z+2x}{2(n+1)} = f_{n+1}$
- $Y^{(opt)} = \frac{n}{n+1} W$, with chunk sizes $\frac{Y^{(opt)}}{n} = \frac{W}{n+1}$

Proof (2/2)

- $E(Y) = W - (z + x)W^2 - (f_n + x)Y^2 + (z + 2x)WY$
- $Y^{(opt)} = \frac{z+2x}{2(f_n+x)} W$
- $E_{opt}(W, n + 1) = E(Y^{(opt)}) = W - \alpha W^2$,
  where $\alpha = z + x - \frac{(z+2x)^2}{4(f_n+x)}$
- By induction, $f_n + x = \frac{(n+1)z+2x}{2n} + x = \frac{(n+1)(z+2x)}{2n}$
- Finally, $\alpha = z + x - \frac{n(z+2x)}{2(n+1)} = \frac{(n+2)z+2x}{2(n+1)} = f_{n+1}$
- $Y^{(opt)} = \frac{n}{n+1} W$, with chunk sizes $\frac{Y^{(opt)}}{n} = \frac{W}{n+1}$

Proof (2/2)

- $E(Y) = W - (\mathrm{z}+\mathrm{x})W^2 - (\mathrm{f}_n+\mathrm{x})Y^2 + (\mathrm{z}+2\mathrm{x})WY$
- $Y^{(\mathrm{opt})} = \frac{\mathrm{z}+2\mathrm{x}}{2(\mathrm{f}_n+\mathrm{x})}W$
- $E_{\mathrm{opt}}(W, n+1) = E(Y^{(\mathrm{opt})}) = W - \alpha W^2$, where $\alpha = \mathrm{z}+\mathrm{x} - \frac{(\mathrm{z}+2\mathrm{x})^2}{4(\mathrm{f}_n+\mathrm{x})}$
- By induction, $\mathrm{f}_n + \mathrm{x} = \frac{(n+1)\mathrm{z}+2\mathrm{x}}{2n} + \mathrm{x} = \frac{(n+1)(\mathrm{z}+2\mathrm{x})}{2n}$
- Finally, $\alpha = \mathrm{z}+\mathrm{x} - \frac{n(\mathrm{z}+2\mathrm{x})}{2(n+1)} = \frac{(n+2)\mathrm{z}+2\mathrm{x}}{2(n+1)} = \mathrm{f}_{n+1}$
- $Y^{(\mathrm{opt})} = \frac{n}{n+1}W$, with chunk sizes $\frac{Y^{(\mathrm{opt})}}{n} = \frac{W}{n+1}$

Outline

Symmetric functions

Definition

Given $n \geq 1$, for $0 \leq i \leq n$, $\sigma_i^{(n)}$ denotes the i-th symmetric function of $x_1, x_2, \ldots, x_n$:

$$\sigma_i^{(n)} = \sum_{1 \leq j_1 < j_2 < \cdots < j_i \leq n} \prod_{k=1}^{i} x_{j_k}.$$

By convention $\sigma_0^{(n)} = 1$

For instance with $n = 3$, $\sigma_1^{(3)} = x_1 + x_2 + x_3$, $\sigma_2^{(3)} = x_1 x_2 + x_1 x_3 + x_2 x_3$ and $\sigma_3^{(3)} = x_1 x_2 x_3$

Optimal solution

Theorem

When z = 0 *(no communication cost), the optimal solution to* DISTRIB(p) *is to send* P_i *a chunk of size* $\frac{\prod_{k \neq i} x_k}{\sigma_{p-1}^{(p)}} W$*, and*

$$E_{opt}(W, p) = W - \frac{\sigma_p^{(p)}}{\sigma_{p-1}^{(p)}} W^2$$

- Closed-form formula ☺ ☺
- Proof by induction

Outline

Optimal solution (1/2)

Theorem

When using the ordering $P_1, P_2, \ldots, P_p$, the optimal solution is to send P_i a chunk of size $\alpha_{i,p} W$, and

$$E_{opt}(W, p) = W - \mathrm{f}_p W^2$$

- *For* $p \geq 1$, $\mathrm{f}_p = \dfrac{\sum_{i=0}^{p} \lambda_i \sigma_{p-i}^{(p)} \mathrm{z}^i}{\sum_{i=0}^{p-1} \lambda_i \sigma_{p-i-1}^{(p)} \mathrm{z}^i}$, *with* $\lambda_i = \frac{4(1+i)}{2^i}$
- $\alpha_{1,1} = 1$, *and for* $p \geq 2$, $\alpha_{p,p} = \dfrac{2\mathrm{f}_{p-1} - \mathrm{z}}{2(\mathrm{f}_{p-1} + \mathrm{x}_p)}$
- $\alpha_{1,p} = 1 - \alpha_{2,p}$ *for* $p \geq 2$
- $\alpha_{i,p} = \dfrac{\mathrm{z} + 2\mathrm{x}_{i-1}}{2(\mathrm{f}_{i-1} + \mathrm{x}_i)}(1 - \alpha_{i+1,p})$ *for* $p > i \geq 2$

Optimal solution (2/2)

Theorem

In the general case, the optimal solution to $\textsc{Distrib}(p)$ *does not depend upon the ordering of the communications from the master*

- Easy algorithm ☺ but no closed-form formula ☹
- Quite complicated proof (still by induction) ☹

Outline

Conclusion

- First extension to master-slave divisible load approach **with unrecoverable failures**
- Nice set of results, similar to classical setting ☺
- Turned out more difficult than expected (☺ or ☹?)
- Tractability of case with different link bandwidths?

Perspectives

- Resources with different risk functions (different owner categories?)
- Case with different speeds, different link bandwidths and different risk functions
- Combine with **replication strategies**
- Combine with multi-round techniques
- Comparison with dynamic approaches