1

情報セキュリティを考慮したコンピュータシステムの最適バックアップ方策

総合経営学科　成瀬　健一郎

背景

転ばぬ先のバックアップ
バックアップの必要性

コンピュータシステムは、いつも記録装置の故障やランサムウェアへの感染の脅威にさらされている。

図1 ランサムウェアに感染した画面
出典:株式会社アクト　https://act1.co.jp/column/0132-2/

図2 病院の電子カルテがランサムウェアによって停止
出典:WIRED
https://www.wired.com/story/universal-health-services-ransomware-attack/

図3 オイルパイプラインが動作を停止
出典:CISA
https://www.cisa.gov/news-events/cybersecurity-dvisories/aa20-049a

システムモデル

バックアップの方法は？

コンピュータシステムが、コントロール

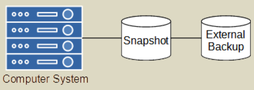

図4 コンピュータシステムと、スナップショットバックアップ、外部バックアップ

1. スナップショットバックアップは、バックアップを一瞬で取得することができるが、ファイルシステムのパフォーマンス低下を及ぼす。ソフトウェア障害が起こったときは回復が出来るが、ハードウェア障害が起こった場合は回復できない。
2. 外部バックアップは、バックアップを取得するのに時間がかかるが、ファイルシステムのパフォーマンスに影響を及ぼさない。また、ハードウェア障害、ソフトウェア障害のどちらが起こっても回復できる。

バックアップのタイミングは？

図5 一定長のタスクとバックアップ計画

1. 実行時間が決まっている1つのタスクを繰り返し実行するシステム
2. 実行時間SをN個に分割し、分割した部分の最後にSnapshot backupを行い、最後にExternal backupを行う。

故障が起こったときのコンピュータの動作は？

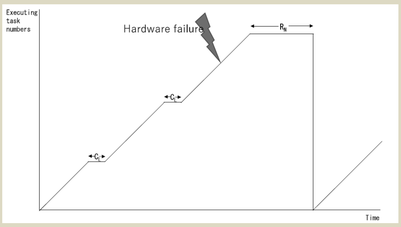

図7 ハードウェア障害が検出された場合、システムは、R_H時間かかって復旧動作を行い、タスクを最初からやり直す。

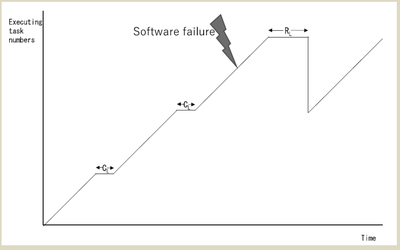

図8 ハードウェア障害が検出された場合、システムは、R_S時間かかって復旧動作を行い、タスクを1つ前からやり直す。

パフォーマンス解析

コンピュータシステムのデータの壊れる確率は？

ハードウェア障害(ハードディスク故障など)
故障する確率は指数分布に従うとすると、累積故障確率は

$$F_H(t) = 1 - \mathrm{e}^{-\lambda_H t}$$

となる。λ_H は、ハードウェア障害による単位時間当たりの平均故障回数。

ソフトウェア障害(ランサムウェア感染など)
故障する確率は指数分布に従うとすると、累積故障確率は

$$F_S(T) = 1 - \mathrm{e}^{-\lambda_S T}$$

となる。 λ_S は、ソフトウェア障害による単位時間当たりの平均故障回数。

ハードウェア障害の平均故障時間$1/\lambda_H$
ソフトウェア障害の平均故障時間$1/\lambda_S$
Snapshot backupを行う時間　C_L
External backupを行う時間　C_N
Snapshot backupを使って回復する時間　R_L
External backupを使って回復する時間　R_N
オリジナルタスク実行時間　S
Snapshot backupによるファイルシステム低下率　w ($0<w<\infty$)

何を求めるのか？

バックアップなどの処理を追加し、故障率からエラーが発生した場合の回復時間を考えた場合、実行時間が決まっているタスクをいくつに分割したら、実行時間が最も短くなるか？

1. バックアップ間隔を短くすれば、障害が起こった場合のやり直し時間が小さくて済むが、タスクの処理以外にかかる時間が多くなって、システムパフォーマンスが悪くなる。
2. バックアップ間隔を長くすれば、障害が起こった場合のやり直し時間が大きくなるが、タスクの処理以外にかかる時間が少なくなり、システムパフォーマンスの悪化が抑えられる。
3. **二律背反(トレードオフ)の状態になる**

パフォーマンス解析(一定長)

どうやって求めるのか？

処理が完了するまでのプロセス $k(k=1,2,\cdots,N)$ の処理時間は、以下の式で表すことができる。

$$\begin{aligned}\widetilde{L}(k) &= C_L + (1+kw)T + \overline{F}_S(T)\overline{F}_H(T)\widetilde{L}(k+1) + F_S(T)\overline{F}_H(T)\left[\widetilde{L}(k)+R_L\right] \\ &\quad + F_H(T)\left[\widetilde{L}(1)+R_N\right] \qquad (k=1,2,\cdots,N-1), \\ \widetilde{L}(N) &= C_L + (1+Nw)T + \overline{F}_S(T)\overline{F}_H(T)C_N + F_S(T)\overline{F}_H(T)\left[\widetilde{L}(N)+R_L\right] \\ &\quad + F_H(T)\left[\widetilde{L}(1)+R_N\right]. \end{aligned} \tag{1}$$

以下のように定義して、

$$\begin{aligned} A &\equiv 1 - F_S(T)\overline{F}_H(T), \qquad B \equiv \overline{F}_S(T)\overline{F}_H(T), \\ C &\equiv C_L + R_L F_S(T)\overline{F}_H(T) + R_N F_H(T), \qquad A - B = F_H(T), \end{aligned}$$

式(1)を解くと、

$$L(N) \equiv \widetilde{L}(1) = \frac{A^{N-1}}{B^N}\sum_{j=1}^{N}\left\{\left(\frac{B}{A}\right)^{j-1}\left[C+(1+jw)T\right]\right\} + C_N \qquad (N=1,2,\cdots). \tag{6}$$

を得る。故障率は、指数分布に従うとし、ソフトウェア障害の発生確率を、$F_S(T) = 1-\mathrm{e}^{-\lambda_S T}$ ハードウェア障害を $F_H(T) = 1-\mathrm{e}^{-\lambda_H T}$ として式(6) に代入すると、平均実行時間である

$$\begin{aligned} L(N) = &\frac{\left[1-\mathrm{e}^{-\lambda_H T}+\mathrm{e}^{-(\lambda_S+\lambda_H)T}\right]^{N-1}}{\mathrm{e}^{-(\lambda_S+\lambda_H)S}}\sum_{j=1}^{N}\left\{\left[\frac{\mathrm{e}^{-(\lambda_S+\lambda_H)T}}{1-\mathrm{e}^{-\lambda_H T}+\mathrm{e}^{-(\lambda_S+\lambda_H)T}}\right]^{j-1}\right. \\ &\left.\times\left[C_L + R_L\left(1-\mathrm{e}^{-\lambda_S T}\right)\mathrm{e}^{-\lambda_H T} + R_N\left(1-\mathrm{e}^{-\lambda_H T}\right) + (1+jw)T\right]\right\} + C_N \\ &(N=1,2,\cdots). \end{aligned} \tag{7}$$

を得る。

数値例を求める

1. $T=S/N$ を式(7)に代入し、N=1とする。
2. $L(N)-L(N+1)$ の計算を行う
3. 2.の結果がマイナスならN に1を足して、2.の計算を行う。
4. 2.の結果がプラスまたは0なら、そのN が実行時間を最も短くなる最適値である。

例として、
S=10，C_L =0.1，C_N=0.4，R_L =0.2，R_N=0.8　として、計算を行うと、表1を得る。

表1 $S=10.0,\ C_L=0.1,\ C_N=0.4,\ \overline{R}_L=0.2\ \ R_N=0.8.$ の時、最適なN_{c^*}と全実行時間$L(N_{c^*})$

		w=0.01		w=0.05		w=0.10	
λ_H	λ_S	N_C^*	$L(N_C^*)$	N_C^*	$L(N_C^*)$	N_C^*	$L(N_C^*)$
0.010	0.050	7	13.261	4	14.715	4	16.209
0.025	0.050	7	14.644	5	16.314	4	18.012
0.050	0.050	8	17.312	6	19.412	4	21.560
0.075	0.050	9	20.517	6	23.102	5	25.722
0.100	0.050	10	24.390	7	27.537	6	30.787
0.050	0.010	5	16.014	4	17.414	3	18.813
0.050	0.025	6	16.556	4	18.233	3	19.962
0.050	0.050	8	17.312	6	19.412	4	21.560
0.050	0.075	10	17.978	7	20.451	5	22.975
0.050	0.100	11	18.583	8	21.408	6	24.287

例えば、$w=0.01,\ \lambda_H=0.050\ \ \lambda_S=0.050.$ の時、最適値であるN_{c^*}は8で、システムの全実行時間$L(8)$は17.312となる。これは、元の実行時間よりも、7.312長くなる。つまり、上記の条件においては、タスクを8つに分割し、分割したところでSnapshot backupを実行した場合が、一番実行時間が短くなる。

パフォーマンス解析(ランダム長)

どうやって求めるのか？

ランダムタスクの平均実行時間は、指数分布に従うとし、$1/\theta \equiv \int_0^\infty \overline{G}(t)\mathrm{d}t$ とする。

式(1)より、

$$\begin{aligned}\widetilde{L}_R(k) &= \int_0^\infty \left\{C_L + (1+kw)t + \overline{F}_S(t)\overline{F}_H(t)\widetilde{L}(k+1) + F_S(t)\overline{F}_H(t)\left[\widetilde{L}(k)+R_L\right]\right. \\ &\quad \left. + F_H(t)\left[\widetilde{L}(1)+R_N\right]\right\}\mathrm{d}G(t), \qquad (k=1,2,\cdots,N-1), \\ \widetilde{L}_R(N) &= \int_0^\infty \left\{C_L + (1+Nw)t + \overline{F}_S(t)\overline{F}_H(t)C_N + F_S(t)\overline{F}_H(t)\left[\widetilde{L}(N)+R_L\right]\right. \\ &\quad \left. + F_H(t)\left[\widetilde{L}(1)+R_N\right]\right\}\mathrm{d}G(t). \end{aligned} \tag{8}$$

が得られる。

$$\begin{aligned} A &= 1 - \int_0^\infty F_S(t)\overline{F}_H(t)\mathrm{d}G(t), \qquad B \equiv \int_0^\infty \overline{F}_S(t)\overline{F}_H(t)\mathrm{d}G(t), \\ C &\equiv C_L + \int_0^\infty \left[R_L F_S(t)\overline{F}_H(t) + R_N F_H(t)\right]\mathrm{d}G(t), \\ A - B &= \int_0^\infty F_H(t)\mathrm{d}G(t). \end{aligned} \tag{9}$$

と置き、式(8)を解くとN タスク分の平均実行時間

$$\widetilde{L}_R(1) = \frac{A^{N-1}}{B^N}\sum_{j=1}^{N}\left[\left(\frac{B}{A}\right)^{j-1}\left(C+\frac{1+jw}{\theta}\right)\right] + C_N. \tag{10}$$

が得られる。また、1回あたりの実行時間は、

$$L_R(N) \equiv \frac{\widetilde{L}_R(1)}{N} = \frac{A^{N-1}}{NB^N}\sum_{j=1}^{N}\left[\left(\frac{B}{A}\right)^{j-1}\left(C+\frac{1+jw}{\theta}\right)\right] + \frac{C_N}{N} \qquad (N=1,2,\cdots). \tag{11}$$

となる。
故障率は指数分布に従うとし、ソフトウェア障害の発生確率を $F_S(T) = 1-\mathrm{e}^{-\lambda_S T}$ 、ハードウェア障害発生確率を $F_H(T) = 1-\mathrm{e}^{-\lambda_H T}$ とする。

式(9)をラプラス・スティルチェス変換をすると、

$$\begin{aligned} A &= 1 - G^*(\lambda_H) + G^*(\lambda_S+\lambda_H), \quad B = G^*(\lambda_S+\lambda_H), \\ C &= C_L + R_L\left[G^*(\lambda_H) - G^*(\lambda_S+\lambda_H)\right] + R_N\left[1-G^*(\lambda_H)\right], \\ X &= \frac{G^*(\lambda_S+\lambda_H)}{1-G^*(\lambda_H)+G^*(\lambda_S+\lambda_H)}, \end{aligned} \tag{14}$$

と、変換できる。但し、$G^*(s) \equiv \int_0^\infty \mathrm{e}^{-st}\mathrm{d}G(t)$。また、$X \equiv B/A$ とする。

また、タスク処理時間 $G(t)$ が、平均値 $1/\theta$の指数分布に従うとすると、$G(t) = 1-\mathrm{e}^{-\theta t}$ となり、式(14)より

$$\begin{aligned} A &= \frac{\lambda_H}{\theta+\lambda_H} + \frac{\theta}{\theta+\lambda_S+\lambda_H}, \qquad B = \frac{\theta}{\theta+\lambda_S+\lambda_H}, \\ C &= C_L + R_L\left(\frac{\theta}{\theta+\lambda_H} - \frac{\theta}{\theta+\lambda_S+\lambda_H}\right) + R_N\frac{\lambda_H}{\theta+\lambda_H}, \\ X &= \frac{\theta}{\theta+\lambda_S+\lambda_H}\Big/\left(\frac{\lambda_H}{\theta+\lambda_H} + \frac{\theta}{\theta+\lambda_S+\lambda_H}\right). \end{aligned} \tag{15}$$

を得る。

数値例を求める

1. 式(15)を式(11)に代入し、N=1とする。
2. $L_R(N)-L_R(N+1)$ の計算を行う
3. 2.の結果がマイナスならN に1を足して、2.の計算を行う。
4. 2.の結果がプラスまたは0なら、そのN が実行時間を最も短くなる最適値である。

例として、
S =10，C_L =0.1，C_N=0.4，R_L =0.2，R_N=0.8　として、計算を行うと、表2を得る。

表2 $S=10.0,\ C_L=0.1,\ C_N=0.4,\ \overline{R}_L=0.2\ \ R_N=0.8.$ の時、最適なN_R^* と1回あたりの実行時間$L(N_R^*)$

		$\theta=1$		$\theta=3$		$\theta=5$	
λ_H	λ_S	N_R^*	$L_R(N_R^*)$	N_R^*	$L_R(N_R^*)$	N_R^*	$L_R(N_R^*)$
0.010	0.050	2	1.565	4	0.641	6	0.451
0.025	0.050	2	1.630	4	0.661	5	0.463
0.050	0.050	2	1.742	4	0.695	5	0.481
0.075	0.050	1	1.822	3	0.721	4	0.499
0.100	0.050	1	1.873	3	0.748	4	0.515
0.050	0.010	2	1.674	4	0.684	5	0.476
0.050	0.025	2	1.699	4	0.688	5	0.478
0.050	0.050	2	1.742	4	0.695	5	0.481
0.050	0.075	2	1.785	3	0.701	5	0.484
0.050	0.100	2	1.828	3	0.708	5	0.487

パフォーマンス解析(ランダム長)

バックアップのタイミングは？

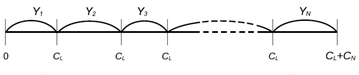

図6 ランダム時間長のタスクとバックアップ計画

1. 実行時間が指数分布に準拠するランダムなタスクを実行するシステム
2. 1つのタスクY_N ($N>0$)を連続実行し、タスクの終了ごとにSnapshot backupを行い、複数回タスクを実行した後External backupを行う。

※処理時間は終わってみないと分からないので等分割できない。

何を求めるのか？

1. ランダムタスクを何回実行してからExternal backupをおこなったら、実行時間が最も短くなるか？

2. **二律背反(トレードオフ)の状態になる**

まとめ

ファイルストレージの記憶容量は増大し、マルウェアへの感染も増加している。何も対策していないコンピュータシステムの大容量ファイルにソフトウェア障害やハードウェア障害が発生した場合、被害は甚大となる。このような事態を防ぐためには、適切なタイミングでファイルストレージのバックアップを作成する必要がある。

本発表では、２種類の障害の発生と２種類のバックアップ方法について考え、バックアップやリカバリの時間を考慮に入れた実行時間が最も短くなる最適なバックアップ回数とその実行時間の求める式を導出し、数値例を提示している。

バックアップポリシーを検討するのに、実行時間が同じであるバッチ処理には一定長のタスクが適しており、様々な長さのタスクがランダムに発生する可能性があるクラウドシステムには、ランダム長タスクが適している。

本手法および結果は、マルウェアやハードウェアの故障により被害を受ける可能性のあるコンピュータシステムのバックアップポリシー作成に有用である。