2005－EVA－12（7）
2005／3／23

ＩＢＭメインフレームでの動的なチャネル構成変更

河野　知行　　株式会社アイ・アイ・エム　東京都文京区本郷２－２７－２０

ＩＢＭのｚシリーズでは,業務実行状況に合わせ最適なシステム構成をOSが判断し,動的なシステム構成の変更を可能としている。例えば,同一プロセッサ内で動作する複数OS間でのCPU能力配分比の自動調整,局所的な入出力要求の偏りにも対応するチャネル構成の自動変更などである。この研究では,z/OSが採用する最適な入出力システム構成の判定手法について整理し報告する。

Dynamic Channel-path Management of IBM z/Series

Tomoyuki Kawano, IIM Corporation, 2-27-20 Hongo Bunkyou-ku, Tokyo

Dynamic Channel-path Management is a new capability, designed to dynamically adjust the channel configuration in response to shifting workload patterns. This is implemented by exploiting new and existing functions in both software and hardware components, and provides the ability to have the system automatically manage the number of ESCON and FICON paths available to supported DASD. This paper describes functions and performance of this z/Series's new facility.

１、はじめに

DCM[1]は業務負荷に応じてチャネル構成を最適化する機能であり,WLM[2],IOS[3],HCD[4],動的I/O再構成機能[5]などのソフトウェア機能,ｚシリーズのCPC[6],ESCON[7]ダイレクタ,ディスク制御装置などのハードウェア機能を組み合わせて実現される。DCMは動的にチャネル構成を変更するが,その対象はディスク制御装置に接続されたESCONまたはFICON[8]ブリッジ経由の光チャネルに限られている。

DCMを使用できるのはｚシリーズのCPCで,z/アーキテクチャモードのz/OSを動作させているシステムである。z/OSのシステム管理機能を提供するWLMには２種類の制御モード（互換モードとゴールモード）がある。WLMに互換モードとゴールモードがあるように,DCMにもバランスモードとゴールモードがある。バランスモードでは,システム全体のスループットが最大化されるようなアクションが取られ,ゴールモードでは重要業務のレスポンスが最適化されるようなアクションが取られる。

WLMは業務プログラムの実行状況を監視し,必要に応じてCPUやメモリ,入出力処理の優先順位制御を行っている。一方,DCMはチャネル構成の最適化を図る。

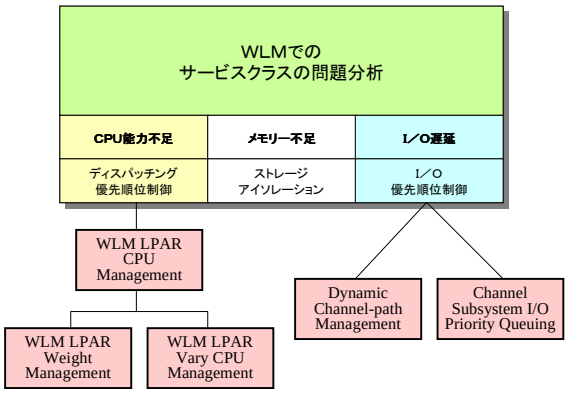


図１　WLMとDCMの関係

２,チャネル本数と負荷バランス

複数のディスク制御装置をシステムに接続する場合,それぞれの制御装置に接続されたディスク装置の稼動状況に応じて,適切なチャネル数を確保する必要がある。注意深くチャネル数を決定したとしても,とある制御装置ではより多くのチャネルを必要としているのに,他の制御装置は必要以上のチャネルを占有すると言うことが往々にして発生する。

システムスループットを最大化するには,全てのチャネルの使用率が均等であることが望まれる。しかし,そのようなシステムは存在しない。また,昼間と夜間の業務特性が違うために,チャネル使用率の偏り具合も,昼間と夜間で大きく変化する。このような事態を放置すれば,負荷が偏ったチャネルの使用率は高くなり,そのチャネルの遅延時間（使用を待たされる時間）は長くなる。結果として,パフォーマンスの悪化を招く。

1 Dynamic Channel-path Management
2 Workload Manager
3 Input Output Supervisor
4 Hardware Configuration Dialog
5 Dynamic I/O Reconfiguration
6 Central Processing Complex
7 Enterprise Systems Connection
8 Fiber Connection

DCM はチャネルの遅延時間を常時監視し,その時点でより多くのチャネルを必要とする制御装置へチャネルを動的に移動させる。このような機能を提供することにより,システム全体のレスポンスを良好に保とうとする。

チャネル使用率が均等化された場合の効果を検証してみよう。図 2 のようなチャネル構成を考える。2台のディスク制御装置に 2 本のチャネルが接続されている。左図では LCU[9]1 に 300 回／秒,LCU2 に 100 回／秒のアクセスが行われる。右図では,LCU1 と LCU2 とも 200 回／秒のアクセスが行われる。両図とも,アクセス回数の総和は同じ 400 回／秒である。

この構成で,全てのアクセスがキャッシュヒットし,5ミリ秒のデータ転送が必要であったとする。その際のチャネルの遅延時間を待ち行列計算式（M/M/2）で求めよう。左図の場合,チャネル遅延時間は LCU1 で 6.43 ミリ秒／アクセス,LCU2 で 0.33 ミリ秒／アクセスとなる。右図の場合,LCU1 と LCU2 とも 1.67 ミリ秒となる。平均チャネル遅延時間を比べると左図で 4.91 ミリ秒／アクセス,右図で 1.67 ミリ秒／アクセスとなる。結果としてチャネル使用率が均等化される右図の方が,チャネル遅延時間が短いことが判る。

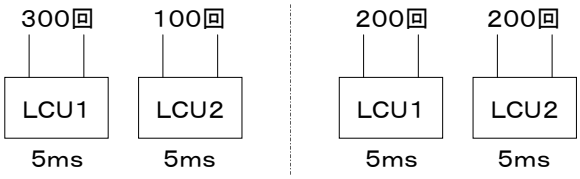


図2　チャネル使用率の均等化の効果

3,チャネル構成定義

DCM は,今までシステム管理者を悩ましていたチャネル構成の最適化を図ってくれる。また,チャネル構成の設計時に考慮しなければならないことを大幅に簡略化してくれる。その点を考察しよう。

先ず,DCM を使用しないシステムで,システム管理者がチャネル構成を検討する際,考慮すべき点を整理する。

- ピーク時間帯の業務負荷を処理するに必要な制御装置のチャネル数を見極める。
- チャネルに複数の制御装置を接続する場合,それぞれの制御装置の稼動状況を時間帯ごとに分析する。
- 可能な限りチャネル使用率をバランスさせるため,制御装置ごとのチャネル数を決定する。
- 設計した構成をハードウェア（HCD）で定義する。
- 実際に運用されている時にパフォーマンスやチャネル使用率の変動状況を監視する。

DCM はチャネル遅延を引き起こしている制御装置へ,自動的にチャネルを移動する。このため,DCM を使用すればチャネル本数に関するのキャパシティ計画やチューニング（負荷バランスの最適化）が不要になる。

DCM は,無制限にチャネルの移動を行う訳ではない。システム管理者が指定した制御装置ごとの最小と最大のチャネル本数の範囲内で,最適なパフォーマンスと最大の可用性を確保するために必要な構成変更を動的に行う。

DCM を使用することにより新たなチャネルを増設することなく,ディスク装置群で処理できるトラフィック量を増やすことが可能になる。また,新たなハードウェアを追加した際にも,その特性に合わせた接続形態を自動的に決定してくれる。

DCM を使用する場合に必要なチャネル構成定義の考慮点を整理しよう。

- ディスク装置群が接続される制御装置単位に DCM の管理対象（DCM 管理制御装置）か否（DCM 非管理制御装置）かを検討する。
- DCM 管理制御装置には,DCM 非管理チャネル（従前と同様のチャネル）を必要最低限,割り当てる。また,DCM が追加することのできる最大チャネル数も決定する。
- DCM 非管理制御装置には,従前と同様に,DCM 非管理チャネル数を決定する。
- システム全体の DCM 管理チャネル数を決定する。

DCM が管理する制御装置のことを DCM 管理制御装置と呼ぶ。また,DCM が動的に構成変更を行う際に使用するチャネルのことを DCM 管理チャネルと呼ぶ。それら以外の制御装置やチャネルの取り扱いは,DCM を使用していない時と同様である。

4,最大チャネル数の制限

z シリーズの CPC でも,最大チャネル数は 256 本に制限されている。システムによっては「256 本以上のチャネルを必要としている」とされるものもある。しかし,実際に 256 本分のチャネル転送速度を必要としているシステムは少ない。また,これはデバイスアドレス制限の問題でもない。多くの場合,256 本のチャネルで提供されるデバイスアドレスの総数は,実際接続しようとしているデバイス数を大きく超えている。

それでは 256 本のチャネルでは足りないとされている問題の本質はどこにあるのか。良好なパフォーマンスを確保するために必要な最大チャネル数を制御装置ごとに割り当てると,256 本以上のチャネルが必要となることである。チャネル転送速度（チャネル使用率）

9　Logical Control Unit

の問題でもなければ,デバイスアドレス（デバイス数）の問題でもない。

DCM は,この問題に焦点を当てていると言える。システム管理者は制御装置に必要なパフォーマンスを確保できるように,チャネルあたりに接続可能なデバイス数を決定する。一方,DCM は業務負荷の変動に合わせ,動的なチャネル数の制御を行う。

例えば,チャネルに接続された制御装置の稼働率が高まれば,DCM はチャネル遅延時間を短縮しようと,現在使用されていないかもしくは使用率が低いチャネルを稼働率が高い制御装置に移す。

DCM は,全ての入出力動作をキャッシュ経由（非同期モード）でおこなうディスク制御装置で,且つ ESCON もしくは FICON ブリッジ（FCV）のチャネル経由で接続されたディスク制御装置のみをサポート対象としている。また,DCM は動的に制御装置のチャネル数を制御するため,その制御装置は複数チャネルが接続できるものでなければならない。

DCM は制御装置のチャネル数を追加したり削減したりするため,該当制御装置はスイッチに接続され,且つスイッチに DCM 管理チャネルが接続されていなければならない。本研究のレポートの中ではスイッチと呼ぶのは ESCON ディレクタのことである。

5,I/O ベロシティ

先にも述べたように,DCM はディスク制御装置のチャネル遅延時間を常時監視する。その結果を示す指標として,制御装置ごとに I/O ベロシティと呼ばれる計算値を算出する。

I/O ベロシティの考え方は,WLM のゴールモードにおける実行ベロシティの考え方に基づいたものである。資源を使用するために必要となる待ち時間の割り合いを数値として把握するために使用する。

$$\text{ベロシティ} = \frac{\text{使用時間}}{\text{使用時間}+\text{待ち時間}} \times 100$$

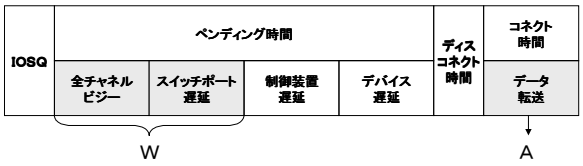
DCM はデータ転送時間（コネクト時間）を使用時間としている。制御装置単位での I/O ベロシティの計算が行われるので,その制御装置に接続されたデバイスのデータ転送時間の合計が,制御装置の総使用時間となる。

待ち時間には,チャネル遅延とスイッチポート遅延が含まれる。この待ち時間を算出するために,DCM はペンディング時間から制御装置遅延時間とデバイス遅延時間を引き算している（図 3 を参照）。ペンディング時間は,入出力操作を司るプロセッサ（SAP[10]）が入出力起動命令（SSCH[11]）を受信した時から,デバイスがチャネルプログラムの実行を受け付けるまでの時間である。ペンディング時間には,次のものが含まれる。

- デバイスをアクセスするためのチャネルが全て使用中である時間（全チャネルビジー）。
- 他のチャネルが制御装置の HA[12]を使用しているためスイッチポートが使用中であり,入出力操作の起動が待たされた時間（スイッチポート遅延）。
- 制御装置の内部パスが使用中であるために入出力操作の起動が待たされた時間。この時間はペンディング時間の制御装置遅延時間として計上される。
- 他のシステムがデバイスを使用中であるためデバイスがリザーブされており入出力操作の起動が待たされた時間。この時間はペンディング時間のデバイス遅延時間として計上される。

SAP が SSCH 要求を受け付けるまでの時間も待ち時間となる。通常,この時間は 0.5 ミリ秒以下である。SAP 待ちとなっている SSCH の待ち行列の長さは,SAP 待ち行列長として報告される。

図 3 に,z/OS 環境での入出力装置のレスポンス時間内訳を示す。ペンディング時間とコネクト時間は,前述のとおりである。IOSQ[13]は先行した入出力操作の完了を待っている時間である。また,ディスコネクト時間は,キャッシュミスが発生しディスク装置をアクセスしている時間である。



$$\text{I／Oベロシティ} = \frac{A}{A+W} \times 100$$

図3　I/O ベロシティの計算

6,ソフトウェアの役割

z/OS でシステム管理機能を提供する WLM は,ディスク装置群のパフォーマンス情報を収集し,入出力操作を司る IOS と共同してディスク制御装置ごとの I/O ベロシティ値を算出する。求められたベロシティ情報は結合機構に格納され,同一 z シリーズの CPC で動作する他のシステムとで共有される。

---

10 System Assist Processor
11 Start Sub Channel
12 Host Adapter
13 Input Output Supervisor Queue

IOS は結合機構に格納された情報を基に,チャネル遅延時間が許容範囲を超えたディスク制御装置を割り出す。この作業のことをバランス検査[14]と呼ぶ。バランス検査で問題があるとされた制御装置が見つかると,そのチャネル遅延時間を許容範囲に収めるアクションが取られる。この処理をバランス補正[15]と呼ぶ。

DCM がゴールモードで動作していれば,WLM が重要な業務に影響を与える制御装置で達成すべき I/O ベロシティを明示的に指示する。I/O ベロシティは I/O レスポンス時間におけるチャネル遅延の影響を測定しており,DCM が構成変更を行うトリガーになるものである。

DCM は次の条件を満たしているシステムで効果を発揮する。

- チャネル使用率が均等でない。また新たなチャネルを追加することなく,現在のチャネルの使用効率を高める必要がある。
- 時間帯により,制御装置への負荷が大きく変動する。
- CPC の 256 本のチャネル制限に近づいている。
- 多くの制御装置を使用しており,一つのチャネルに複数の制御装置を接続している。

DCM は制御装置へ最大のチャネルを接続し,全てのチャネルが一つの制御装置にしか接続されていない場合には無力である。しかし,DCM が可用性の向上に役立つ機能も提供している。可用性の向上は本研究の対象ではないため,ここでは割愛する。

7,バランスモードでの制御

IPL[16]での初期化が完了すると,DCM は DCM 管理制御装置へのチャネルの追加と移動の制御を行うことが可能となる。DCM は新たにチャネル追加を必要とする DCM 管理制御装置を検出するために,バランスモードとゴールモードの 2 種類のアルゴリズムを準備している。

WLM の動作モード（互換モードやゴールモード）に関わりなく,DCM をバランスモードで動作させることができる。このモードでは,全ての DCM 管理制御装置の I/O ベロシティを同じ範囲に収めようとする。

DCM のもう一つの制御方式であるゴールモードを選択するには,WLM がゴールモードで動作していることが必要である。WLM は目標を達成していないサービスクラス（業務グループ）を見付けると,その遅延理由を調査する。例えば,CPU 待ち,ページング待ち,入出力操作待ち,チャネル遅延などに分類し,その待ち要因を一つ一つ吟味する。

重要なサービスクラスでチャネル遅延が大きな割り合いを占めると,WLM は DCM の助けが必要と判断する。すると WLM は,チャネル遅延を生じさせている制御装置を割り出し,その目標 I/O ベロシティを更新する。つまり,このモードではサービス目標を達成していない業務を遅延させている制御装置のスループットを向上させようとする。

I/O ベロシティは,DCM 管理制御装置に DCM 管理チャネルを追加すべきか否かを判断するための基準として使用される。バランスモードで,この I/O ベロシティを算出する際の手法の詳細を紹介しよう。

インターバルごとに,WLM は IOS に I/O ベロシティ計算に必要なデータ収集を指示する。この場合,全てのディスク制御装置がデータ収集の対象となる。これは DCM 管理制御装置が非管理制御装置と同じチャネルを共有することがあるためである。DCM が管理チャネルを移動させると,他の制御装置のチャネル使用率が増加し,パフォーマンスの悪化を招く場合がある。これはドミノ効果とも呼ばれるものである。DCM は一つの変更が及ぼす影響を予測し,このドミノ効果による悪影響を排除する。

IOS はディスク制御装置ごとの CMB[17]やサブチャネル情報を収集し,制御装置ごとに集約する。これらの情報は,このシステムから見た制御装置の稼動情報であり,同一プロセッサ内で稼動するシステムごとにデータ収集が行われる。IOS はインターバルごとに,これらの差分を算出し WLM へ通知する。

同一プロセッサ内で複数システムが動作している場合,WLM は結合機構を使用して,他のシステムとのデータ集約を図る。つまり,それぞれのシステムで収集されたデータは,結合機構上で集約され更新される。集約された I/O ベロシティ値は,WLM により IOS に通知され,制御装置ごとに管理される。バランスモードでは,この情報を基に DCM 管理制御装置の目標 I/O ベロシティを設定し,DCM 管理チャネルの追加・移動を判断する。

通信し合うシステムの一つでも結合機構へのアクセスが行えなくなると,全てのシステムが結合機構をアクセスできるようになるまで,DCM は構成変更を行わない。これは DCM が誤った情報で構成を変更しないようにするためである。

例えば,本番システムとテストシステムがあるとしよう。本番システムが一時的に結合機構をを使用できない間,テストシステムのためだけに構成変更が行われることは回避したい。この問題を解決するには,全てのシステムが結合機構をアクセスできていることを確認する必要がある。

---

14 Balance Checking
15 Imbalance Correction
16 Initial Program Load

17 Channel Measurement Block

8,ゴールモード処理

バランスモードは,全ての DCM 管理制御装置の I/O ベロシティを同じ範囲に収めようとするものであった。ここでは,もう一つ制御方式であるゴールモードについて紹介しよう。

ゴールモードを選択するには,WLM がゴールモードで動作していることが必要である。WLM がゴールモードで動作している場合,10 分ごとにポリシー調整ルーチンが実行される。このポリシー調整ルーチンは,その時点でサービス目標を達成していないサービスクラスを捜し出す。目標を達成していないサービスクラスを見つけると,その遅延理由を調査し,必要なアクションを取る。

遅延理由がチャネル競合による遅延である場合,WLM は遅延を引き起こしている DCM 管理制御装置を割り出す。最初に最も遅延を引き起こしている制御装置の目標ベロシティを向上させた場合の効果予測を行う。その結果,効果ありと判断された場合は,その DCM 管理制御装置の目標ベロシティが更新される。予測の結果,効果が無いと判断された場合,次に多くの遅延を引き起こしている制御装置の予測が行われる。このようにして,遅延を解消するために目標ベロシティを向上させるべき DCM 管理制御装置が検索される。

サービスクラスには,そのサービス目標（パフォーマンス目標）を持っていないシステム系のサービスクラスがある。これらのサービスクラスについても,WLM が遅延要因におけるチャネル遅延の割り合いを調査し,チャネル遅延が多いと判断した場合に前述と同様の処理を行う。

WLM は,DCM 管理制御装置のベロシティを改善するために DCM を起動するが,DCM はポリシー調整ルーチンの一部として実行されるものではない。また,DCM は一回の起動により一つの制御装置のベロシティ改善しか行えない訳でもない。DCM は一回の起動で同一プロセッサ内で稼動する複数システムからの DCM 管理制御装置のベロシティ調整要求を処理する。このため,該当システムのサービスクラスでチャネル遅延が発生しなくとも,そのシステムで DCM が起動されることにより,DCM 管理制御装置のベロシティ調整が行われることがある。

DCM が DCM 管理制御装置のベロシティ調整を行った場合,他のシステムでは,それ以降の数インターバルの間,DCM 管理制御装置のベロシティの調整を行わないようにしている。これは I/O ベロシティ値が低い場合の過剰反応を防止するためである。

目標ベロシティを持つ DCM 管理制御装置のベロシティ改善が行われる前に,その制御装置を使用しているサービスクラスのチャネル遅延による遅延が解消された場合,WLM はその制御装置の目標ベロシティを省略値に変更する。この時点で,その DCM 管理制御装置のベロシティ調整は行われなくなる。これは,急激な構成変更を予防するための処置である。

WLM は,DCM 管理制御装置の目標ベロシティを設定するだけであり,実際にチャネルの追加や移動の必要性を判定するのは DCM の一つの機能として IOS が行うインバランス調整機能である。つまり,WLM は IOS と協調して,IOS が最適な構成変更方策の選択を行っているのみである。

システムの制御モードがゴールモードから互換モードに切り替えられた場合,DCM はゴールモードからバランスモードに切り替わる。その際,DCM 管理制御装置の目標ベロシティは一旦無効にされる。

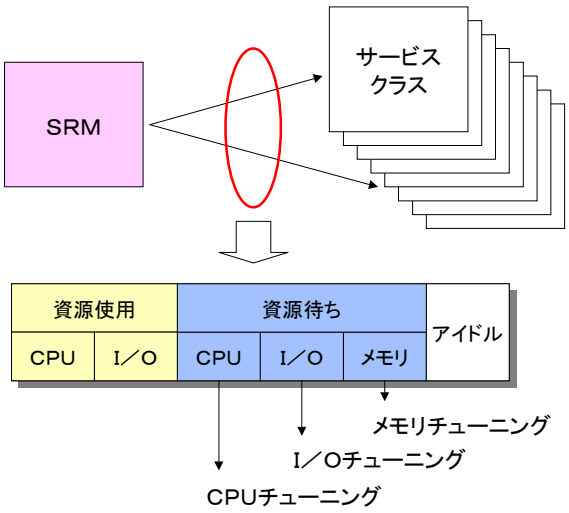


図４　WLM のポリシー調整ルーチンの判断

9,バランス確認

同一プロセッサ内で稼動するシステムは,バランスモード,ゴールモードの如何に関わらず,必要に応じて DCM 管理制御装置の目標ベロシティを設定する。指定された目標ベロシティを基にチャネルの追加や移動の必要性を判定し,チャネルの移動を実行するのが DCM のバランス確認,インバランス調整機能である。これらの機能は IOS の一部として実行される。

このバランス確認とインバランス調整では,次の４つのことが行われる。

- DCM 管理制御装置に２つのチャネルが接続されていることを保証する。チャネルが一つしか準備されていない場合,障害や間違ったオペレーションでチャネルがオフラインにされた場合に問題が生じるためである。
- WLM により設定された（ゴールモードの場合の）目標ベロシティを達成していない DCM 管理制御装置が無いこと。これはバランス確認の機能である。

・バランスモードで判定された目標ベロシティを達成していない DCM 管理制御装置が無いこと。これもバランス確認の機能である。
・達成されていない目標ベロシティを持つ DCM 管理制御装置のベロシティを改善するための制御を行う。これはインバランス調整の機能である。

１０,インバランス調整

バランス確認では,最初に省略値のベロシティを達成していない DCM 管理チャネルを捜し出す。省略値以上のもの,省略値以下のものの両方が選択される。また同時に,ゴールモードで設定された目標ベロシティを達成していない DCM 論理制御装置も検索対象となる。但し,この際,最近チャネルの追加や移動を行った DCM 管理制御装置は除外される。

ゴールモードで動作している場合,IOS は選択された DCM 管理制御装置の情報を WLM に通知し,WLM から優先してベロシティ調整を行うべき制御装置を知らせてもらう。IOS は,優先順位順に DCM 制御装置のベロシティを改善するために必要なアクションを決定する。その際,次のような点を考慮する。

・ダイレクタポートビジーが遅延原因ではないか。もしそうであれば,現存のチャネルを他のチャネル（例えば未使用のものや,同じスイッチであるが他のポートのもの）に切りかえる。
・チャネルを追加すればベロシティを向上させることができるか。もしそうであれば,どのチャネルを追加すべきか。
・チャネルを追加する場合,デバイスは今以上のチャネルをサポートしているか。例えば,一つのシステムから制御装置へのチャネル数が８本を超えないか。もしくは制御装置へのチャネル数が,システム管理者が作成した構成定義（HCD）で許された以上の本数にならないか。
・DCM 管理制御装置へチャネルを追加する場合,そのチャネルに既に接続されている他の制御装置への影響は。
・DCM 管理制御装置に接続されたチャネルに接続されている他の DCM 管理制御装置を切り離すと,この DCM 管理制御装置のベロシティは改善されるか。
・DCM 管理制御装置に接続されたチャネルに接続されている他の DCM 管理制御装置を切り離すと,その DCM 管理制御装置のベロシティは悪化しないか。
・チャネルの移動を行う場合,その移動により障害発生時に閉塞状態となる可能性があるデバイスが無いか。
・チャネルの追加を行う場合,そのチャネルは他の制御装置で使用されていないか。
・ゴールモードの目標ベルシティを持つ DCM 管理制御装置のベロシティを改善するためのアクションが,ゴールモードの目標ベロシティを持つ DCM 管理制御装置に悪影響を与え場合,IOS は WLM にその旨を通知し,より重要なサービスクラスに影響を与えないかを確認する。

これらの条件がクリアーされた時点で,DCM は初めて判断を下す。チャネルの追加や移動を行う際,その構成変更を行った時刻が記録され,その後,数インターバルの間,その DCM 管理制御装置への構成変更が行われないようにする。この構成変更が禁止されている間に,構成変更の効果を反映した新しいベロシティ値が算出されるはずである。

１１,参考文献

SG24-5952 z/OS Intelligent Resource Director
(IBM's RedBooks)