บทที่ 2

การเลือกตัวพารามิเตอร์และรูปแบบทางคณิตศาสตร์ที่ใช้ในการวิเคราะห์

จุดประสงค์ของบทนี้ คือ ให้ผู้ศึกษามีความรู้ความเข้าใจเกี่ยวกับตัวพารามิเตอร์สำหรับการประเมินสมรรถนะของระบบคอมพิวเตอร์ และรูปแบบทางคณิตศาสตร์ที่ใช้ในการประเมิน เพื่อประโยชน์ต่อการวิเคราะห์สมรรถนะของระบบต่อไป

สิ่งที่จะได้รับจากบทนี้

1) แนวความคิด (concept) ในการเลือกตัวพารามิเตอร์ที่เหมาะสมเพื่อใช้ในการ
2) รูปแบบทางคณิตศาสตร์เชิงวิเคราะห์
3) รูปแบบแถวคอย (Queuing Modeling)
4) วิธีการเลือกและกำหนดตัวแปร

วิธีการหรือเทคนิคที่ใช้ในการประเมินสมรรถนะของระบบคอมพิวเตอร์ที่สำคัญประกอบด้วย

1) การใช้บุคคลเป็นผู้ตรวจสอบ
2) การใช้ข้อมูลซึ่งอธิบายการทำงานของระบบหรือข้อมูลด้านบัญชีของระบบ
3) การใช้ตัวเฝ้าตรวจด้านซอฟต์แวร์
4) การใช้ตัวเฝ้าตรวจด้านฮาร์ดแวร์
5) การคิดเปรียบเทียบสมรรถนะ
6) การวิเคราะห์ด้วยแบบจำลอง
7) การวิเคราะห์ด้วยรูปแบบทางคณิตศาสตร์

การใช้บุคคลเป็นผู้ตรวจสอบ การใช้บุคคลเป็นผู้สังเกตสมรรถนะของระบบคอมพิวเตอร์โดยไม่ต้องใช้ข้อมูลที่ซับซ้อนหรือใช้เครื่องมือในการวิเคราะห์ที่ยุ่งยาก เป็นการกระทำที่ดี การเฝ้าสังเกตโดยตรงเช่นนี้ทำให้ได้ประเมินเรื่องที่สำคัญต่าง ๆ ของระบบคอมพิวเตอร์ไปด้วย เช่น ห้องคอมพิวเตอร์ควรมีสภาพแวดล้อมที่เหมาะสมมีอุณหภูมิหรือความชื้นพอเหมาะ ซึ่งต้องมีการตรวจเช็คเสมอ นอกจากนี้ยังมีตัวแปรของสภาพแวดล้อมที่วัดไม่ได้อีก เช่น การดูแลรักษาความสะอาด ฝุ่นหรือสิ่งสกปรกซึ่งอาจทำให้สมรรถนะของเทปหรือหน่วยจานแม่เหล็กเสียหาย ทำให้ผลการปฏิบัติงานเสียหายไปด้วย บุคคลซึ่งมีหน้าที่วิศวกรระบบต้องรู้จักพิจารณาว่าสิ่งใดควรทำ สิ่งใดก่อให้เกิดผลกระทบต่อระบบ การสื่อสารระหว่าง

พนักงานผู้ควบคุมเครื่องอาจมีผลกระทบต่อสมรรถนะของระบบได้ พนักงานต้องรู้จักสังเกตและเอาใจใส่ต่อการทำงานของอุปกรณ์ต่างๆ

ส่วนขอบเขตที่ใช้ในการศึกษาสมรรถนะของระบบ ได้แก่ เวลาครบวงงานและความเร็วในการตอบสนอง การที่วิศวกรระบบสามารถปรับการใช้ระบบคอมพิวเตอร์ได้ตามที่ผู้ใช้ต้องการเป็นการเพิ่มสมรรถนะให้ระบบคอมพิวเตอร์ซึ่งบางครั้งข้อมูลของระบบยังไม่ปรากฏออกมา วิธีที่ดีสำหรับการตรวจวัดเวลาครบวงงานคือ การส่งงานซึ่งประมวลผลแบบกลุ่มให้ระบบประมวลผลและรอผลลัพธ์ สำหรับการตรวจวัดความเร็วในการตอบสนองคือ การใช้คำสั่งเพื่อทำงานในหน้าที่ต่างๆ กัน

การใช้ข้อมูลด้านบัญชีของระบบ ในวงการคอมพิวเตอร์นั้นผู้ใช้ต้องมีความรู้เกี่ยวกับระบบคอมพิวเตอร์ก่อนจึงจะมีความเข้าใจในข้อมูลที่เกี่ยวข้องกับสมรรถนะของระบบไม่เช่นนั้นอาจเกิดความเข้าใจผิดได้ ข้อมูลด้านบัญชีของระบบเป็นปัญหาที่สำคัญในระบบคอมพิวเตอร์ซึ่งใช้งานหลายด้าน การทำงานของระบบมัลติโปรแกรมมิ่งนั้นเวลาที่งานชิ้นหนึ่งใช้หน่วยความจำหลัก หน่วยความจำสำรอง และเวลาที่ใช้ในการประมวลผลจะมีนัยยะสำคัญแตกต่างกันไปแม้ว่าจะประมวลผลงานชิ้นเดียวกัน ทั้งนี้ขึ้นกับว่างานอื่นซึ่งปฏิบัติการอยู่ในขณะเดียวกันนั้นใช้ทรัพยากรอะไรของระบบบ้าง จึงไม่สามารถยืนยันได้ว่าข้อมูลด้านบัญชีของระบบจะยุติธรรมสำหรับผู้ใช้ทุกราย แต่ระบบปฏิบัติการทุกระบบจะบรรจุด้วยระบบย่อยของข้อมูลด้านบัญชีที่ทำไว้อย่างละเอียด โดยจะเก็บรายละเอียดของรายการเกี่ยวกับการใช้ประโยชน์จากทรัพยากรต่างๆของระบบที่งานนั้นใช้ ข้อมูลที่ได้จากการปฏิบัติงานประจำวัน ปัจจัยหนึ่งซึ่งสำคัญในการวัดสมรรถนะของระบบคือปริมาณงานในขบวนการผลิตที่ระบบปฏิบัติการจริงประจำวัน รายงานผลการปฏิบัติงานประจำวันของระบบช่วยให้สามารถรวบรวมจำนวนตัวพารามิเตอร์ ด้านสมรรถนะและนำมาใช้เป็นข้อพึงระมัดระวังถึงสิ่งผิดปกติที่อาจเกิดขึ้น

การใช้ตัวเฝ้าตรวจด้านซอฟต์แวร์ ข้อมูลที่จัดเก็บโดยโปรแกรมของระบบนั้นบางครั้งอาจมีรายละเอียดไม่เพียงพอ ดังนั้นกรณีที่ต้องการเก็บข้อมูลเพิ่มขึ้นจึงมีการนำซอฟต์แวร์ช่วยในการจัดเก็บมาใช้โดยซอฟต์แวร์จะช่วยรวบรวมข้อมูลระหว่างที่กระบวนการของงานต่างๆประมวลผลรวมทั้งส่วนที่ระบบประมวลผลด้วย เช่น ซอฟต์แวร์โพรบ (Probes) ซึ่งแบ่งเป็น 2 ประเภท คือ แซมพลิ่งโพรบ (Sampling Probes) และ เทรสซิ่งโพรบ (Tracing Probes)

แซมพลิ่งโพรบ จะเก็บค่าตัวแปรที่น่าสนใจแบบสุ่มเพื่อใช้วิเคราะห์สมรรถนะตัวอย่างข้อมูลที่เก็บนี้สามารถประมาณแบบการกระจายของข้อมูล ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐานได้ จากนั้นจึงวิเคราะห์โดยใช้ทฤษฎีการสุ่มตัวอย่าง

เทรสซิ่งโพรบ เป็นกลไกสำหรับเก็บรายการในรายละเอียดเมื่อมีฟังก์ชั่นงานที่สำคัญถูกกระทำการโดยหลีกเลี่ยงปัญหาต่างๆ เช่น เครื่องทำงานช้าลงหรือจะเกิดโสหุ้ยมากอันเกี่ยวเนื่องจากการเทรส

การใช้ตัวเฝ้าตรวจด้านฮาร์ดแวร์ เครื่องคอมพิวเตอร์ในปัจจุบันจะมีแผงควบคุมไฟฟ้าใช้ในการบ่งชี้เพื่อการควบคุมฟลิปฟล็อปต่าง ๆ ในระบบถูกกำหนด และเมื่อองค์ประกอบต่าง ๆ ของระบบถูกใช้งาน

ในช่วงเวลาที่ฟลิปฟล็อปถูกกำหนดว่า จะเป็นการเตรียมช่วงเวลาการวัดโดยตรงซึ่งฟังก์ชันควบคุมถูกใช้การสังเกตโดยตรงสามารถให้ประมาณการเวลาหยาบ ๆ ทั้งนี้การวัดให้ถูกต้องทำโดยใช้เครื่องมือทางอีเล็คทรอนิคติดไว้ในจุดที่ต้องการทดสอบตามความเหมาะสมในระบบ และรวบรวมข้อมูลเกี่ยวกับสภาพของระบบต่างนั้นสามารถใช้วัดสัดส่วนและเวลากำหนดที่ฟลิปฟล็อปถูกกำหนดว่าเช่นนี้คือพื้นฐานของอุปกรณ์เฝ้าตรวจด้านฮาร์ดแวร์

ตัวพารามิเตอร์หรือตัวบ่งชี้สมรรถนะของระบบ

การประเมินสมรรถนะของระบบขึ้นอยู่กับแนวความคิดของผู้ประเมิน ดังนั้นตัวพารามิเตอร์ที่นำมาใช้ประเมินสมรรถนะจะแตกต่างไปด้วย ตัวพารามิเตอร์เป็นตัววัดที่สำคัญบางตัวกำหนดยาก เช่น ความง่ายในการใช้ระบบ ความเป็นโครงสร้างของโปรแกรมและชุดคำสั่งที่มีเพาว์เวอร์ เป็นต้น ดังตัวอย่าง ตัวบ่งชี้สมรรถนะของระบบคอมพิวเตอร์ ซึ่งทำงานแบบโต้ตอบตามตารางที่ 2.1 ประกอบด้วยรายการต่อไปนี้

ชั้นของตัวแปร	ตัวอย่างตัวแปร	คำจำกัดความทั่วไป
ผลผลิตที่ได้รับ	อัตราปริมาณงานต่อหน่วยเวลา อัตราผลผลิตต่อหน่วยเวลา อัตราความจุหรือปริมาณงานสูงสุด อัตราการกระทำการของชุดคำสั่ง อัตราการประมวลผลข้อมูล	ปริมาณข้อมูลซึ่งระบบประมวลผลในหนึ่งหน่วยเวลา
ความสามารถในการโต้ตอบ	เวลาตอบสนอง เวลาครบรอบวงงาน เวลาที่ใช้ในการตอบกลับ	เวลาระหว่างการนำข้อมูลเข้าสู่ระบบและได้ผลลัพธ์ที่สอดคล้อง
อรรถประโยชน์	มอดุลด้านฮาร์ดแวร์เช่น ซีพียู หน่วยความจำ ช่องไอโอ อรรถประโยชน์ของระบบปฏิบัติการ มอดุลของซอฟต์แวร์ที่ใช้งานร่วมกัน อรรถประโยชน์ของระบบฐานข้อมูล	สัดส่วนระหว่างเวลาซึ่งแต่ละส่วนของระบบใช้งานในช่วงเวลาที่กำหนด

ตารางที่ 2.1 ตัวแปรสำหรับการประเมินสมรรถนะ

จากตารางที่ 2.1 สคมภ์คำจำกัดความทั่วไปบอกให้ทราบว่า ตัวแปรด้านผลผลิตที่ได้รับมีขนาดเท่ากับปริมาณคูณด้วยเวลายกกำลังลบหนึ่ง ตัวแปรด้านความสามารถในการโต้ตอบมีขนาดเท่ากับเวลาที่ใช้ ตัวแปรด้านอรรถประโยชน์ไม่สามารถวัดขนาดได้ ดังนั้น ปริมาณจากการวัดที่แตกต่างกันเกิดจากความแตกต่างของระบบคอมพิวเตอร์และภาระงานของระบบ ทั้งนี้ปริมาณการวัดที่นิยมใช้กันได้แก่ งาน โปรแกรม กระบวนการ ขั้นตอนงาน ภารกิจ รายการเคลื่อนไหว การกระทำที่มีผลกระทบซึ่งกันและกัน และชุดคำสั่ง ซึ่งตัวแปรที่ใช้ทั้งหมดมีความเกี่ยวพันธ์กัน

<u>รูปแบบทางคณิตศาสตร์เชิงวิเคราะห์</u> (Analytic Model)

ในการศึกษาครั้งนี้มีวิธีการเลือกพารามิเตอร์และกำหนดวิธีการประเมินโดยใช้รูปแบบจำลองเชิงวิเคราะห์ (Analytic Model) ซึ่งแบ่งออกเป็น 2 ลักษณะ คือ

1. <u>รูปแบบดีเทอร์มินิสติค</u> (Deterministic Model) เป็นรูปแบบที่แสดงถึงระบบคอมพิวเตอร์ต่าง ๆ ของภาระงานในแต่ละหน่วยงานคอมพิวเตอร์พิจารณาและวิเคราะห์ถึงคุณสมบัติของขั้นตอนวิธีซึ่งมีการจัดการทรัพยากรที่เหมาะสม สำหรับการประมวลผลแบบมัลติโปรเซสเซอร์ (Optimal Schedule Multiprocessor System) และ การจัดลำดับของการใช้หน่วยความจำสำรอง (Secondary Storage Device) โดยใช้รูปแบบของค่าเฉลี่ย (Mean Value Models) โดยพิจารณาค่าเฉลี่ยของภาระงานของงาน ซึ่งนำเข้าสู่ระบบด้วยค่าเฉลี่ยคงที่ (arrival rate) รูปแบบค่าเฉลี่ยสำหรับภาระงานประกอบด้วยกลุ่มของงาน (Job) ที่มีลักษณะเหมือนกัน โดยมีคุณสมบัติของตัวแปรของภาระงานมาประกอบเป็นงานในรูปแบบเหมือนกันซึ่งจะเสมือนมีค่าเฉลี่ยคงที่เป็นค่าเฉลี่ยของการเข้ามาในระบบด้วยค่าอัตราคงที่ (mean arrival rate) ซึ่งจะใช้เป็นค่าสำหรับการทำนายค่าดัชนีคุณภาพและความสามารถของระบบคอมพิวเตอร์

การพิจารณารูปแบบของหน่วยประมวลผลกลาง และหน่วยอินพุทเอาท์พุทที่เหลื่อมกัน โดยมีวัตถุประสงค์เพื่อประเมินและเปรียบเทียบรูปแบบการเหลื่อมกันระหว่างหน่วยประมวลผลกลางและหน่วยอินพุท ในกิจกรรมของงานเดียวกัน ซึ่งมีการกำหนดพารามิเตอร์ ดังตัวอย่างดังนี้

รายการ	สัญญาลักษณ์	ค่า
ขนาดของหน่วยความจำ	M	2k, 4k, 6k, ไบต์
ขนาดของระเบียน	r	64 ไบต์
ขนาดของบล็อก	b	2k ไบต์
เวลาที่หน่วยประมวลผลกลางประมวลผล 1 ระเบียน	t_{cpu}	1.6 ms (milli second)
ช่วงเวลาเข้าถึงหน่วยความจำสำรอง	t_a	15 ms
เวลาที่ถ่ายโอนข้อมูลในหน่วยความจำสำรอง	t_w	10μs (micro second)
จำนวนพาร์ติชั่นในหน่วยความจำหลัก	v	1, 2, 3,

ตารางที่ 2.2 แสดงรายการพารามิเตอร์ สัญญลักษณ์และค่าของหน่วยประมวลผลกลางและหน่วยอินพุทเอาท์พุทที่เหลื่อมกัน

กำหนด g คือ blocking factor ขนาดของบล็อกหารด้วยขนาดของระเบียน

$$g = b/r$$

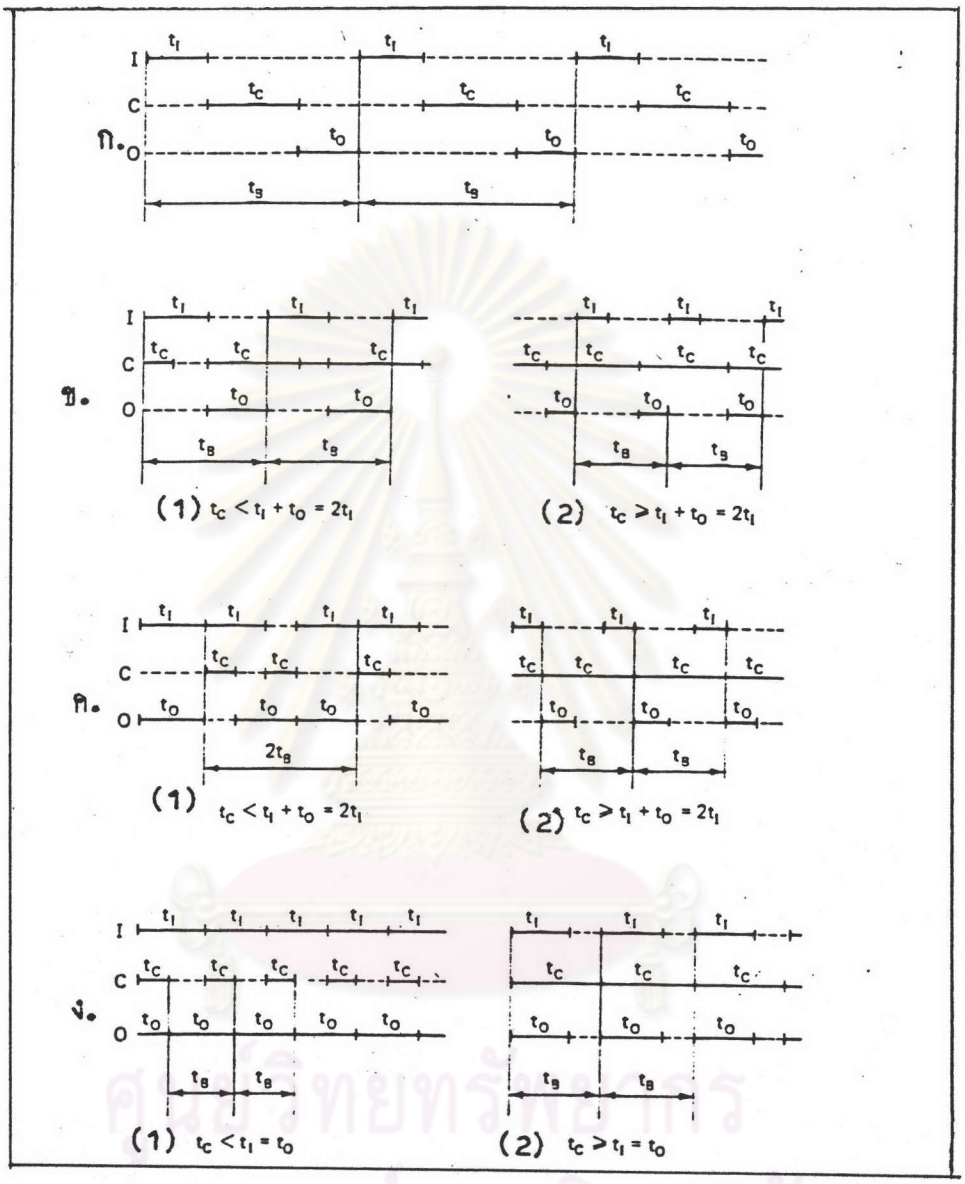
หน่วยความจำหลักถูกแบ่งเป็นพาร์ติชั่นจำนวน v พาร์ติชั่นซึ่งแต่ละพาร์ติชั่นมีขนาดเท่ากับขนาดของบล็อก

$$v = M/b$$

ลักษณะ	ลำดับการเหลื่อมกันคือขนาดพาร์ติชั่น
1. ตามลำดับ (ไม่มีการเหลื่อมกัน)	1
2. สองทางเฉพาะ (C-I หรือ O-C)	2
3. สองทางขนาน (C-I หรือ Q-C หรือ I-O)	3
4. สามทางขนาน (I-C-O)	4

ตารางที่ 2.3 ลักษณะของการประมวลผลเหลื่อมกันสำหรับหน่วยประมวลผลกลางและหน่วยอินพุทเอาท์พุท (I/O)

ถ้าพาร์ติชั่นในหน่วยความจำหลัก (v) = 1 บล็อกต่อไปไม่สามารถโอนเข้าในหน่วยความจำหลักได้จนกว่าบล็อก่อนหน้าจะถูกข้ายออกไปไว้ในหน่วยความจำสำรองดังรูปที่ 2.1 (ก) เวลาที่ใช้ในการนำข้อมูลเข้าประมวลผลและแสดงผลจะได้



รูปที่ 2.1 ผังแสดงเวลาในการทำงานเหลื่อมกันของเวลาคำนวณเวลาอินพุทและเวลาเอาท์พุท

เวลาของอินพุท $t_I = t_a + bt_w$

เวลาของการประมวลผล $t_C = g\ t_{cpu}$

เวลาของเอาท์พุท $t_O = t_a + bt_w = t_I$

เมื่อให้เวลาของการประมวลผลระเบียนหนึ่ง เป็นเวลา 1/T
เวลาที่การประมวลผลบล็อกหนึ่ง เป็นเวลา t_B
ด้านนี้อัตราของทรุพุทในหนึ่งบล็อกต่อหนึ่งหน่วยเวลา เป็น

$$1/T = t_B / g$$

ดังนั้นถ้ามี g ระเบียนในแต่ละบล็อก

$$t_B = t_I + t_C + t_o$$

$$t_B = gt_{cpu} + 2t_a + 2bt_w$$

ถ้าพาร์ติชั่นในหน่วยความจำหลัก (v) = 2 และเป็นแบบสองทางเฉพาะคือ ให้เกิดการทำงานของอินพุทเอาท์พุท และพร้อมกันแต่เฉพาะซีพียูกับอินพุทหรือเอาท์พุทกับซีพียู เท่านั้น โดยทำอินพุทและเอาท์พุทพร้อมกันไม่ได้

ถ้าเกิดเหตุการณ์ไอ/โอบาวด์ (I/O Bound) หมายถึง การใช้เวลาสำหรับหน่วยไอ/โอมากกว่าซีพียู ดังรูปที่ 2.1 (ข.1)

$$t_C < t_I + t_0 = 2\ t_I$$

ถ้าเกิดเหตุการณ์ซีพียูบาวด์ (CPU Bound) หมายถึงการใช้เวลาสำหรับซีพียูมากกว่าหน่วยไอ/โอ ดังรูปที่ 2.1 (ข.2)

$$t_C > t_I + t_0 = 2\ t_I$$

ส่วนถ้าพาร์ติชั่นในหน่วยความจำหลัก (v) = 2 และสองทางขนานกันจะได้ลักษณะที่คล้ายกัน ดังรูปที่ 2.1 (ค.1, ค.2)

ส่วนถ้า v = 3 จะเกิดได้ 2 ลักษณะคือ

$t_C < t_I + t_0$ รูปที่ 2.1 (ง.1)

$t_C < t_I + t_0$ รูปที่ 2.1 (ง.2)

รูปที่ 2.1	สมการ	
ก.	$1/T = t_{cpu} + 2t_e/g + 2rt_w$	$z = 1 + 2y + 2x$
ข.1	$1/T = 2t_e/g + 2rt_w$	$z = 4y + 2x$
ข.2	$1/T = t_{cpu}$	$z = 1$
ค.1	$1/T = t_{cpu}/2 + t_e/g + rt_w$	$z = 1/2 + 2y + X$
ค.2	$1/T = t_{cpu}$	$z = 1$
ง.1	$1/T = t_e/g + rt_w$	$z = 3y + x$
ง.2	$1/T = t_{cpu}$	$z = 1$

ตารางที่ 2.4 แสดงความสัมพันธ์ในรูปสมการ

ในกรณีที่การเหลื่อมระหว่างซีพียูและไอโอสมบูรณ์ ค่าอรรถประโยชน์ของซีพียูเท่ากับ 1 ดังนั้น เวลาที่ซีพียูใช้ประมวลผลหนึ่งเรคคอร์ดได้อัตราส่วน $z = 1/Tt_{cpu}$ มีค่าเท่ากับ 1

รูปแบบ	กำหนด b คงที่ ค่า M ตัวแปร		กำหนดค่า M คงที่ (2K) ค่า b ตัวแปร	
	T (ระเบียน/วินาที)	ρ	T(ระเบียน/วินาที)	ρ
รูปที่ 2.1 (ก.)	261.9	0.419	261.9	0.419
รูปที่ 2.1 (ข.1)	450.9	0.721	317.2	0.507
รูปที่ 2.1 (ค.1)	523.9	0.838	420.6	0.672
รูปที่ 2.1 (ง.2)	625	1.000	488.7	0.781

ตารางที่ 2.5 แสดงการกำหนดค่าคงที่และตัวแปรตามรูปที่ 2.1 และตารางที่ 2.4

เมื่อกำหนดให้ T คือปริมาณงานต่อหน่วยเวลา และ ρ คือค่าอรรถประโยชน์ของซีพียูนั่นคือ

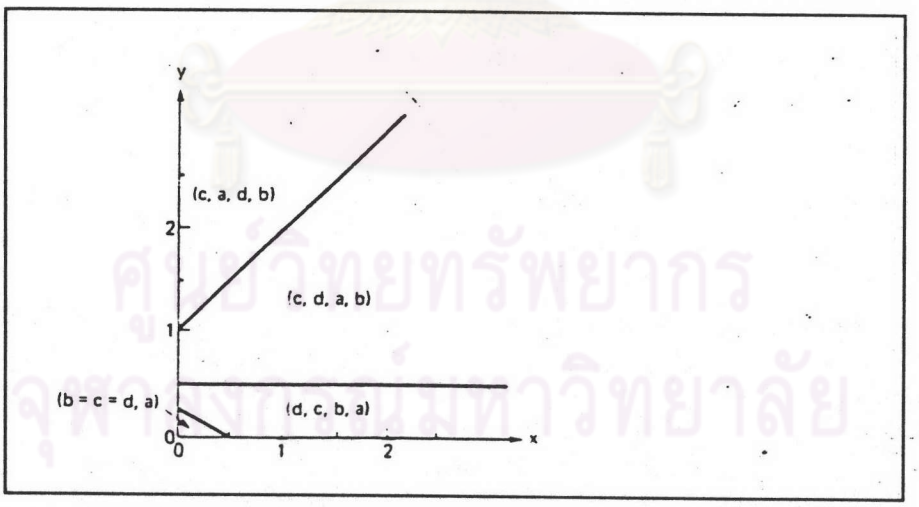
$$\rho = T/T_M = 625 \text{ ระเบียนต่อวินาที}$$

ในกรณีที่ไอโอบาวด์ z มีค่ามากกว่า 1 กำหนดให้ ρ เป็นค่าดัชนีของความสามารถและคุณภาพโดยกำหนดค่าพารามิเตอร์ในรูปแบบเป็นค่าอรรถประโยชน์ของซีพียู ซึ่งสามารถกำหนดสมการโดยมีพารามิเตอร์ 2 ตัวได้แก่

$$x = rt_w / t_{cpu}$$

$$y = rt_a / Mt_{cpu}$$

แทนค่า X, Y จากค่าดังกล่าวเรียงลำดับของการกำหนดวิธีการที่ ง. จะดีที่สุดเรียงตามมาด้วย ก, ข และ ก. ดังรูปที่ 2.2



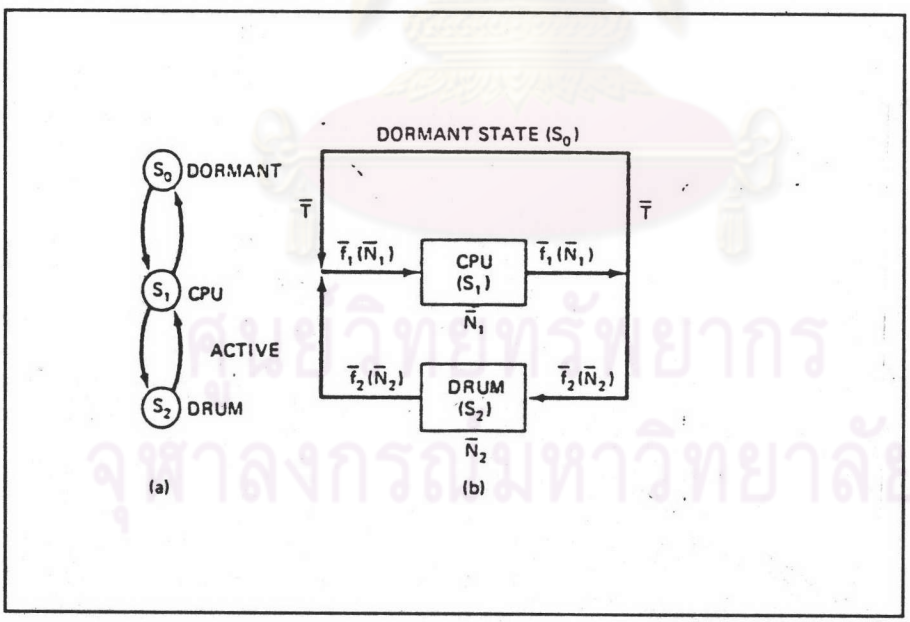
รูปที่ 2.2 แสดงการแบ่งเขตของการลำดับความสามารถและคุณภาพที่เหลื่อมกัน

ค่าเฉลี่ยของระบบมัลติโปรแกรมมิ่ง เพื่อประมาณค่าปริมาณงานต่อหน่วยเวลาโดยมีสมมุติฐานว่าการประมวลผลกิจกรรมของซีพียูและไอโอไม่เหลื่อมกันจากรูปที่ 2.3 แสดงถึงการทำงานแบบงานหนึ่งงาน (ก) เป็นแผนภูมิสถานะของงานหนึ่งงาน ซึ่ง S_0 คือสถานะที่งานไม่อยู่ในหน่วยความจำหลัก S_1

คือ สถานะ รอบริการจากซีพียู S_2 คือสถานะรอบริการจากหน่วยความจำสำรอง (ข) เป็นรูปแบบของระบบ ซึ่งมีหน่วยความจำสำรองเพียงหนึ่งหน่วย และรูปแสดงถึงระบบมัลติโปรแกรมมิง ซึ่งกำหนดให้งาน (JOB) ทุก ๆ งานเหมือนกัน การใช้เวลาซีพียูเป็น t_{cpu} มีจำนวนการเรียกใช้หน่วยความจำสำรองเท่ากับการเรียกใช้เกิดในช่วงเวลาที่สังเกต n_{dr} และเมื่องานประมวลผลด้วยซีพียูเป็นเวลา $t_{cpu}/(n_{dr} + 1)$ จนจะจบงานให้ $\bar{N}_1$ เป็นจำนวนงานโดยเฉลี่ยของสถานะ S_1 และให้ $\bar{N}_2$ เป็นจำนวนงานโดยเฉลี่ยของสถานะ S_2

ดังนั้น $\bar{N}_1 + \bar{N}_2 = N$

ค่าเฉลี่ยของงานที่เข้าระบบจะต้องเท่ากับค่าเฉลี่ยของทั้งที่ซีพียูและหน่วยความจำสำรอง งานที่ออกจากระบบโดยเฉลี่ยขึ้นอยู่กับงานที่บรรจุในสถานะ $\bar{N}_1$ และ $\bar{N}_1$ กำหนดได้คือ $\bar{f}_1(\bar{N}_1)$ และ $\bar{f}_2(\bar{N}_2)$ เมื่อแต่ละงานสู่สถานะ S_1 จำนวน $n_{dr}+1$ ครั้งแรก งานก็จะออกจากสถานะ S_1 จำนวน n_{dr} ครั้งและเข้าสถานะ S_2 และครั้งสุดท้ายงานจะไปที่สถานะ S_0 กำหนดให้ T คือตัวบ่งชี้บอกสมรรถนะของระบบงาน ซึ่งหมายถึงค่าเฉลี่ยของปริมาณงานต่อหน่วยเวลา



รูปที่ 2.3 แสดงรูปแบบของมิลติโปรแกรมมิง และสภาพของการทำงานจะได้อัตราการทำงานสูงสุด (Maximum througput rate)

$$\bar{f}_2(\bar{N}_2) = \frac{n_{dr}}{n_{dr}+1}\bar{f}_1(\bar{N}_2)$$

$$\bar{T} = \bar{f}_1(\bar{N}_1) - \bar{f}_2(\bar{N}_2) = \frac{\bar{f}_1(\bar{N}_1)}{n_{dr}+1}$$

$\bar{f}_1(\bar{N}_1) = 0$ เมื่อ $\bar{N}_1 = 0$

$= \frac{n_{dr}+1}{t_{CPU}}$ เมื่อ $\bar{N}_1 c \geq 1$

$= \bar{N}_1 \frac{n_{dr}+1}{t_{CPU}}$ เมื่อ $0 < \bar{N}_1 < 1$

ในกรณีที่ $\bar{N}_1 = 0$ งานทั้งหมดที่กระทำการจะรอหน่วยความจำสำรอง ซึ่งจะใช้เวลาไม่รู้จบเพื่อให้บริการ นั่นคือ หน่วยความจำสำรองทำงานเต็มเปี่ยม และมีค่าเฉลี่ยของปริมาณงานต่อหน่วยเวลาของระบบเท่ากับ 0 เมื่อ $\bar{N}_1 > 1$ ซีพียูจะไม่ว่างและจะได้อัตราการทำงานสูงสุด (Maxinum throughput rate)

$$T_{max} = \frac{1}{t_{CPU}}$$

$$f_{2max} = f_2(\alpha) = \frac{s}{R_{dr}}$$

$f_2 = 0$ for $\bar{N}_2 = N$

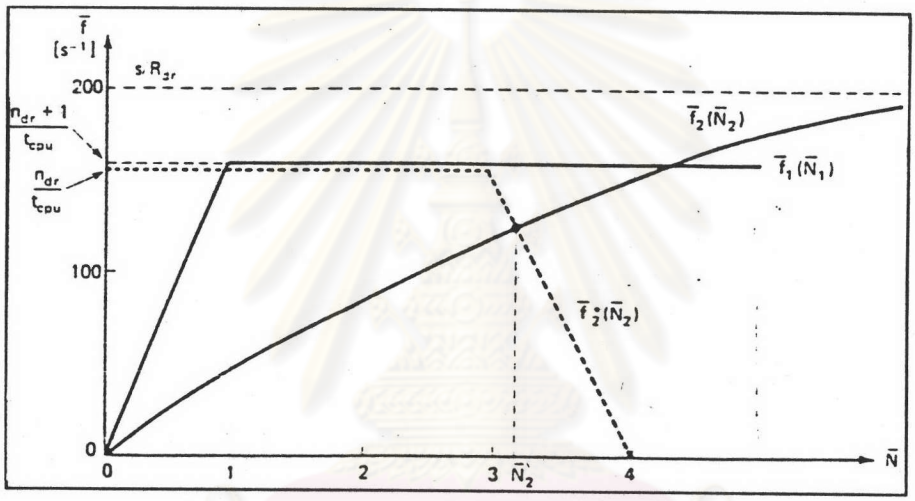
$= \frac{n_{dr}}{t_{CPU}}$ for $\bar{N}_2 \leq N-1$

$$= \frac{n_{dr}}{t_{CPU}} (N - N_2) \text{ for } N - 1 < \bar{N}_2 < N$$

n_{dr}

t_{CPU}

ดังนั้น ในการเพิ่มทรูพุทก็เป็นการลดเวลา t_{CPU} เพิ่มเวลา t_{CPU} ก็ลดทรูพุทและสำหรับ N_{dr} จำนวนเรียกของหน่วยความจำสำรองต่อเวลา t_{CPU} ก็เป็นสัดส่วนกลับกับทรูพุท ดังรูปที่ 2.4



รูปที่ 2.4 แสดงอัตราค่าเฉลี่ยการเข้ารับและออกจากการบริการจากรูปที่ 2.3

2. <u>รูปแบบความน่าจะเป็น</u> (Probabilistics) เป็นรูปแบบที่มีพื้นฐานทฤษฎีการรอคอย (Queue Theory) ซึ่งมีหลักการทางคณิตศาสตร์ ที่จะพิจารณาการประเมินสมรรถนะของระบบคอมพิวเตอร์ ประกอบด้วย

2.1 รูปแบบหรือการแจกแจงของเวลาการใช้บริการ (ServiceTime Distributions) ในรูปแบบของการใช้ทรัพยากรของระบบคอมพิวเตอร์ เช่น มีทรัพยากรอันหนึ่งคือตัวหน่วยประมวลผล เวลาที่โปรแกรมใช้บริการของหน่วยนี้จะประกอบไปด้วย เวลาของการทำงานตามคำสั่งของโปรแกรมและเวลาทั้งหมดที่เสียสำหรับส่วนไอโอ ต้องการใช้จากโปรแกรมตลอดเวลาการเรียกใช้หน่วยความจำเพราะโปรแกรมเรียกใช้หน่วยความจำแบบเสมือน แต่ถ้าทรัพยากรหน่วยนั้นเป็นฐานแม่เหล็กที่

เป็นแบบหัวเคลื่อนที่ เวลาที่โปรแกรมใช้บริการประกอบด้วยเวลาที่ไปหาตำแหน่งของข้อมูล และเวลาที่เปลี่ยนแปลงข้อมูล ระยะทางของข้อมูลถูกเคลื่อนที่ ซึ่งขึ้นอยู่กับความเร็วของหน่วยฐานแม่เหล็กและปริมาณของข้อมูลที่ถูกเคลื่อนที่โดยการเรียกใช้จากโปรแกรมทุก ๆ ส่วนประกอบที่เราจะมาพิจารณากันเราได้จากการวัดตรวจข้อมูล และ โดยการวิเคราะห์ตามรูปแบบของทรัพยากร โดยปกติในแง่ของการใช้งานและทางปฏิบัติแล้ว ลักษณะของเวลาการใช้บริการของทรัพยากรต่าง ๆ นั้นมีลักษณะเป็นแบบสุ่ม ซึ่งรูปแบบที่เราจะมาพิจารณาก็คือ การแจกแจงของเวลาที่ใช้ของทรัพยากรที่เหมาะสม ถ้าเวลาการใช้บริการของทรัพยากรต่าง ๆ ทำการจัดส่วนประกอบอื่น ๆ ตามเวลาที่เกิดขึ้นแบบสุ่ม คือ ไม่สามารถพยากรณ์ไว้ล่วงหน้า แต่เวลาที่ใช้บริการของทรัพยากรนั้น ไม่สามารถกำหนดไว้ล่วงหน้าได้หรือเป็นแบบสุ่ม แต่ก็ไม่ใช่ว่าเราไม่สามารถกำหนดลักษณะของทรัพยากรได้ เช่น เราไม่ทราบล่วงหน้าว่าจะใช้เวลากับทรัพยากรหนึ่ง ๆ เท่าใดได้ แต่เราพอจะทราบได้เป็นโดยเฉลี่ยของเวลาที่ใช้บริการ ลักษณะของค่าเฉลี่ยของเวลาที่ใช้กับหน่วยหนึ่ง ๆ ของทรัพยากรนั้น เราก็นำมาเป็นแนวทางกำหนดรูปแบบได้

การกำหนดรูปแบบในรายละเอียด เราต้องทราบการแจกแจงความน่าจะเป็นของการใช้บริการโดยตัวแปรต่าง ๆ เป็นอิสระต่อกัน คือ เวลาการใช้บริการในหน่วยทรัพยากรต่าง ๆ เป็นอิสระต่อกัน เช่น การใช้บริการของหน่วยอินพุท และเอาท์พุทไม่ได้ขึ้นกับเวลาที่ใช้บริการของหน่วยความจำประมวลผลกลาง เป็นต้น

เมื่อเรากำหนด P(x) เป็น ความน่าจะเป็นของ x

จะได้ว่า $0 \leq P(x) \leq 1$ ทุก ๆ ค่าของ x

และ $\Sigma_x P(x) = 1$

แต่โดยปกติแล้ว ไม่เป็นงานที่จะกำหนดงานด้วยการแจกแจงของรูปแบบนั้นได้โดยตรง แต่เราก็สามารถกำหนดค่าง่าย ๆ ที่ใช้พิจารณาด้วยค่าเฉลี่ยของการแจกแจงแบบนี้ได้ โดยให้ E(x) เป็นค่าเฉลี่ย

$$E(x) = \Sigma_x xP(x)$$

และบางครั้งเราอาจจะเลือก moments และค่าเฉลี่ย moments เป็นค่าเฉลี่ยหรือค่าที่คาดหมายของการใช้เวลาในกำลังที่ n^{th} เช่น

$$E(x^n) = \Sigma_x x^n P(x)$$

เมื่อค่าเฉลี่ย moment ที่ n ก็คือ ค่าคาดหมายของผลต่างของเวลาที่ใช้บริการกับค่าเฉลี่ย เช่น

$$E(x-E\,x)^n = \Sigma_x(x-E\,x)^n P(x)$$

ค่าเฉลี่ย moment ที่ 2 ก็คือค่าแปรผันของการแจกแจงของเวลาที่ใช้บริการ เมื่อถอดรากที่สอง ของค่าแปรผันก็คือ ค่าความเบี่ยงเบนมาตรฐาน (δx)

$$\delta^2 x = E(x^2) - (E(x))^2$$

ค่าเฉลี่ยของเวลาการใช้บริการเป็นตัวกำหนดค่าเฉลี่ยของการใช้เวลาต่อทรัพยากรหนึ่ง ๆ ค่าแปรผัน คือ ค่ากำหนดค่าการแปรปรวนของค่าเฉลี่ยจะมีค่าที่วัดความแปรปรวนกับค่าเฉลี่ยของการใช้บริการเป็นค่าสัมประสิทธิ์ของความแปรปรวน เช่น

$$Cx = \frac{\delta x}{E[x]}$$

เวลาการใช้บริการของหน่วยประมวลผลโดยปกติจะมีค่าความแปรปรวนสูง ค่าของสัมประ - สิทธิ์ของความแปรปรวนจะมีค่าเท่ากับ 10 หรือมากกว่าที่ไม่แปลก แต่หน่วยอินพุทจะมีค่าแปรปรวนต่ำ ค่าสัมประสิทธิ์ต่ำด้วยปกติจะต่ำกว่า 1 แต่ถ้าค่าสัมประสิทธิ์ความแปรปรวนเป็นศูนย์ก็หมายถึงทุก ๆ ค่าเวลาการใช้บริการเท่ากันทุก ๆ ครั้ง

ถ้าตัวแปรของเวลาการใช้บริการมีการกระจายเป็นแบบค่าคงที่ (Discrete) โดยมีการแจกแจงความน่าจะเป็นแบบเรขาคณิต (Geometric Distribution) เช่น กำหนด $O < P < 1$ เป็น

$$P(x) = (1-P)^{x-1} \quad P\,;\, x = 1, 2, 3,...$$

จะมีค่าเฉลี่ยเป็น

$$E(x) = \frac{1}{P}$$

และ

$$E(x)^2 = \frac{2-P}{P^2}$$

ได้ค่าความเบี่ยงเบนมาตรฐาน $\delta x = \frac{\sqrt{1-P}}{P}$

สัมประสิทธิ์ความแปรปรวน $Cx = 1-p$

ถ้าตัวแปรของข้อมูลที่เรานำมาพิจารณาลักษณะของเวลาใช้บริการเป็นแบบต่อเนื่อง (Continuous) จะมีตัวกำหนด 2 ส่วน คือ Probability Density function ($f_x(x_0)$) และ probability distribution function $F_x(X_0)$

Probability density function กำหนดเป็น

$$f_x(X_0) = \frac{dF_x(X_0)}{dx_0}$$

$$F_x(X_0) = \int_{-\infty}^{0} x f_x(X_0) dx_0$$

และ $f_x(x_0)dx_0 = 1$

จากข้อกำหนดการหาค่าเฉลี่ย โมเมนต์ (moment) และค่าแปรปรวน ดังนี้

$$E(X) = \int_{-\infty}^{\infty} x f_x(X_0) dx_0$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_x(X_0) dx_0$$

และ $$E(X^2) = \int_{-\infty}^{\infty} (x - E(x))^2 f_x(X_0) dx_0$$

ถ้าการกระจายของตัวแปรเป็นการกระจายความน่าจะเป็นแบบยูนิฟอร์ม (uniform)

$$f_x(x_0) = \frac{1}{b-a}, a \le x_0 \le b$$

$$0,$$

$$f_x(x_0) = \begin{cases} 0 & , x_0 < a \\ \dfrac{x_0 - a}{b-a} & , a < x0 < b \\ 1 & , x_0 > b \end{cases}$$

$$E(x) = \frac{a + b}{2}$$

$$E(x^2) = \frac{a^2 + ab + b^2}{3}$$

$$\delta x^2 = \frac{(b-a)^2}{12}$$

และ $$Cx = \frac{b-a}{(b+a)\sqrt{3}}$$

ดังนั้น เราจะได้สมการของความน่าจะเป็นในช่วง $a < x \leq b$ เช่น

$$p[a < x \leq b] = \int_a^b f_x(x_0)dx_0$$

หรือ $$p[a < x \leq b] = F_x(b) - F_x(a)$$

การแจกแจงความน่าจะเป็นอีกแบบหนึ่งที่นิยมใช้คือ การแจกแจงความน่าจะเป็น แบบเอกซโพเนนเชียล (Exponential Distribution) โดยมีอัตราการให้บริการ = a

$$f_x(x_0) = \begin{cases} 0 & , x_0\ 0 \\ ae^{-ax0} & , x_0\ 0 \\ 0 & , x_0\ 0 \end{cases}$$

$$f_x(x_0) = 1 - ae^{-ax0} \quad , x_0\ 0$$

$$E\,x = \frac{1}{a}$$

$$E\,x^2 = \frac{2}{a^2}$$

$$\delta x^2 = \frac{1}{a^2}$$

และ $Cx = 1$

การนำการแจกแจงความน่าจะเป็นแบบเอกซ์โพเนนเซียล (Exponential Distribution) มาใช้กับรูปแบบความสัมพันธ์ทางคณิตศาสตร์ของระบบคอมพิวเตอร์ โดยแบ่งลักษณะของการแจกแจงน่าจะเป็นสองแบบตามความแปรปรวน ถ้าการแจกแจงความน่าจะเป็นของตัวแปร ที่มีการแปรปรวนมาก ที่จะใช้การแจกแจงความน่าจะเป็นแบบไฮเปอร์เอกซ์โพเนนเซียล (Hyper exponential) ส่วนการแจกแจงความน่าจะเป็นของตัวแปรที่มีค่าการแปรปรวนน้อย ก็จะเลือกใช้แบบไฮโปเอกซ์โพเนนเซียล (Hypo exponential)

3. <u>หลักการจากจัดลำดับงานเพื่อเข้ารับบริการ</u> (Scheduling Algorithm)

นอกเหนือจากค่าเฉลี่ยของเวลาที่ใช้บริการของทรัพยากรหน่วยต่าง ๆ ของระบบแล้ว ลักษณะที่สำคัญเกี่ยวกับทรัพยากรของระบบยังมีอีกสิ่งหนึ่ง คือ การจัดลำดับงาน หรือโปรแกรมเพื่อเข้ารับบริการจากทรัพยากรของระบบในหน่วยต่าง ๆ เป็นหลักการที่จะตัดสินใจว่า ทรัพยากรจะถูกจัดให้ทำงานโปรแกรมส่วนใดและให้สำเร็จได้หรือเรียกอีกอย่างหนึ่งว่าการจัดคิวของโปรแกรมให้ไปให้ทรัพยากรส่วนต่าง ๆ ของระบบ

หลักในการจัดงานเพื่อเข้ารับบริการ มีหลายวิธีการดังนี้ คือ

การจัดแบบงานใดเข้าก่อนก็รับบริการก่อน (First-Come First-Served = FCFS) สำหรับแบบ FCFS นี้ไม่เหมาะสำหรับใช้กับทรัพยากรที่มีความแปรปรวนของเวลาให้บริการสูง สำหรับการจัดแบบนี้งานที่รับบริการจะถูกจัดตามเวลาที่งานนั้นเข้ามารับบริการ งานใดที่เข้ามาก่อนก็รับบริการก่อนตามเวลาที่เข้ามาไม่มีการแซงคิว

การจัดแบบงานใดเข้ามาทีหลังก็รับบริการก่อน (Last Come First Served = LCFS) แบบที่ตรงข้ามกับแบบแรกทุก ๆ งานที่เข้ามาทีหลังก็จะถูกจัดให้รับบริการก่อนเป็นการลัดคิวหรือแซงคิว งานที่เข้ามาก่อน ก็เกิดปัญหาว่าจะแซงคิว งานที่รับบริการอยู่แล้วให้หยุดก่อน เอางานที่เพิ่งเข้ามาใหม่ไปใช้บริการหรือจะแซงคิวงานที่จะถูกเลือกให้รับบริการงานถัดไป

การจัดแบบงานใดเข้ามาทีหลังก็รับบริการก่อนโดยการแซงคิว (Last Come First Served Preemptive Resume = LCFSFR) คือการจัดที่ต้องแซงคิวทุกครั้งที่เขามารับบริการ ซึ่งเป็นลักษณะของการให้บริการของหน่วยประมวลผลในระบบอินเทอร์แอกทิฟซ์ (Interactive System)

โดยทั่ว ๆ ไป การจัดงานเพื่อรับบริการของหน่วยประมวลผลจะจัดเป็นแบบบราวด์โรบิน (Round-Robin = LRR) หรือบางทีเรียกว่า ไทม์สไลซิงค์ (Time-Slic) ซึ่งเป็นวิธีการที่กำหนดว่าในช่วงของเวลาซึ่งเรียกว่า ควีนตัม(Quantum) หรือ ไทม์สไลซ์ (Time-Slic) งานที่รับบริการจะถูกจัดแบบงานใดเข้ามาก่อนรับบริการก่อน (FCFS) ตลอดเวลาภายในเวลาที่กำหนด แต่ถ้างานใดทำงานในช่วงเวลาควีนตัม นั้นยังไม่เสร็จงานนั้นจะถูกแซงคิวด้วยงานอื่น แล้วงานนั้นจะต้องไปต่อท้ายของคิวงานที่จะเข้ารับบริการใหม่

เฉพาะฉะนั้น เวลาที่งานหนึ่ง ๆจะเข้ารับบริการจากทรัพยากรนี้ต้องแบ่งเวลาเป็นหลาย ๆ ควันตัม ซึ่งจะทำให้หน่วยประมวลผลได้ถูกใช้ร่วมกันจากหลาย ๆ งาน อย่างมีประสิทธิภาพมากขึ้นงานที่ใช้เวลาในการบริการสั้นก็ไม่จำเป็นต้องถูกบังคับให้รองานที่ใช้เวลาการบริการก่อนหน้านี้เสร็จไปก่อน แต่การที่จะเปลี่ยนจากงานที่ให้บริการอยู่ไปให้งานใหม่นั้นก็เสียเวลาไปบ้าง ซึ่งปกติจะน้อยมากเป็น 100 ไมโครวินาที (100 ของเศษหนึ่งส่วนพันวินาที) ต่อการแซงต่อหนึ่งครั้ง ดังนั้นการเลือกช่วงของควันตัมให้มากกว่าเวลาที่เสียไปต่อการเปลี่ยนงานอันเกิดจากการแซงคิวของงานแต่ถ้าควันตัมเป็นเวลามาก ๆ ก็เสมือนเป็นแบบ FCFS ซึ่งทำให้การใช้ทรัพยากรนั้นมีประสิทธิภาพน้อยลง

แต่ถ้าเรามาคำนึงถึงให้มีค่าเฉลี่ยของเวลาการถามตอบข้อมูลน้อยที่สุด(Mean Response Time) การจัดแบบใช้เวลาสั้นเข้าทำงานก่อน (Short-Remaining-Time-First) (SRTF) งานที่เหลือเวลาในการรับบริการน้อยที่สุดทำก่อน คืองานที่เข้ามาใหม่มีเวลาการใช้บริการทรัพยากรนั้นน้อยกว่างานที่รับบริการอยู่ งานที่เข้ามาใหม่จะถูกแซงคิวให้ได้ทำงานทันที ลักษณะนี้เหมาะสมสำหรับส่วนอินพุทและเอาท์พุท เช่น ของดรัมแม่เหล็กแซงคิวด้วยเวลาที่ลาเทนซีสั้นที่สุดทำงานก่อน (Shortest-Latency-Time-First) ส่วนของฐานแม่เหล็กที่เป็นเวลาที่ซีคสั้นที่สุดทำงานก่อน (Shortest-seek-time-first)

ที่กล่าวมาแล้วยังไม่คำนึงถึงกรณีที่มีการให้โอกาสเปลี่ยนแปลงการเข้ารับบริการจากภายนอก (External Priority)

<u>ความสามารถของระบบคอมพิวเตอร์กับความสัมพันธ์ของการแจกแจงความน่าจะเป็นของเวลาการให้บริการกับการจัดงานเข้ารับบริการ</u>

เมื่อพิจารณาถึงความสามารถของระบบจะทำการวัดค่าเฉลี่ยของจำนวนงานที่งานเข้ามารับบริการของระบบต่อหน่วยของเวลา เรียกว่า ทรูพุท(Throughput) ถ้าค่าทรูพุทมากที่สุด (Maximum throughput) ในรูปแบบที่เรากำหนดก็คือค่าสูงสุดของค่าอรรถประโยชน์ของทรัพยากรนั้น ตลอดจนถ้าค่าเฉลี่ยการถามตอบข้อมูลน้อยที่สุด (Minimum Mean Response Time) ก็คือภาระงานยังคงสามารถดำเนินการต่อไปได้

ถ้าเวลาให้บริการมีการแปรปรวนมากคือ $Cx > 1$ งานส่วนใหญ่ใช้เวลาในการบริการมากกว่าค่าเฉลี่ยของเวลาใช้บริการ ซึ่งจะทำให้ความสามารถลดลงเมื่อเราเลือกให้งานเข้ารับบริการแบบ FCFS เพราะงานส่วนใหญ่จะไปรอที่หน่วยประมวลผล ทำให้ส่วนอินพุทและเอาท์พุทมีเวลาการรอคอยให้บริการมาก ทำให้ความสามารถของการทำงานแบบมัลติโปรแกรมมิ่งลดลง แต่ถ้าเป็นแบบราวด์โรบินงานที่ใช้เวลานั้นก็ไม่ต้องเสียเวลาให้งานที่ใช้เวลามากทำงานเสร็จ ก็สามารถแซงคิวเข้าไปรับบริการได้

ความสัมพันธ์ระหว่างการวัดความสามารถของระบบคอมพิวเตอร์

ทรูพุทค่าอรรถประโยชน์และเวลาเฉลี่ยของการให้บริการของระบบคอมพิวเตอร์ (Throughput, Utilization and Mean Service Time) ถ้ามีทรัพยากรหน่วยหนึ่ง และกำหนดค่าเฉลี่ยของเวลาที่ให้บริการหนึ่งงานเป็นค่า E(X) และอัตราเฉลี่ยของการให้บริการเป็น 1/E(x) เราจะเรียกค่าอรรถประโยชน์เมื่ออัตราของเวลาที่ทรัพยากรนั้นทำงานว่า U เพราะฉะนั้น

$$\text{ค่าทรูพุท} = \frac{U}{E(x)}$$

ในกรณีที่มีทรัพยากรเหมือนกัน K หน่วย จะได้ค่า

$$\text{ทรูพุท} = \frac{KU}{E(x)}$$

ทรูพุทค่าเฉลี่ยความยาวของการรอคอย ค่าเฉลี่ยของเวลาการถามตอบข้อมูล(Throughput, Mean Queue lenght, Mean Responsetime)

เมื่อกำหนดค่าเฉลี่ยความยาวของการรอคอย คือ จำนวนงานที่รอที่จะใช้บริการของทรัพยากรหนึ่ง ๆ และเวลารอคอยหรือเวลาที่ถามตอบข้อมูลคือเวลาระหว่างงานเข้ามายังทรัพยากรจนสิ้นสุดการใช้ทรัพยากรนั้น

$$L = E(1) = \sum_{L=1}^{\infty} 1P(1)$$

P(1) : ความน่าจะเป็นของความยาวของการรอคอย 1 เพราะฉะนั้นเวลาเฉลี่ยของการรอคอย (Q)

$$Q = E(g) = \int_{0}^{\infty} f_g(g_0)dg_0$$

ซึ่งจะได้ค่าเป็น L = Q เป็นทรูพุท

สมมติเมื่อกำหนดเวลาเริ่มต้น ตั้งแต่ 0 ถึง T ให้ n(T) เป็นลำดับของงานที่เสร็จสิ้นเมื่อเวลา T

$$\lambda = \frac{n(T)}{T}$$

A(T) คือพื้นที่ของ 1(f) ตั้งแต่ 0 ถึงเวลา T

$$A(T) = \int_0^T 1(t)dt$$

$$L = \frac{A(T)}{T}$$

$$Q = \frac{A(T)}{nT}$$

$$L = \frac{A(T)}{T} = \frac{n(T)}{T} / \frac{A(T)}{n(T)} = Q$$

L : ความยาวเฉลี่ยของการรอคอยเพื่อรับการบริการจากทรัพยากรหนึ่ง รวมงานที่กำลังรับบริการด้วย

Q : ค่าเฉลี่ยของเวลาที่รอคอยเพื่อรับบริการรวมเวลาที่รับบริการด้วย

ปัญหาที่สำคัญของการจัดงานเพื่อเข้ารับบริการมีอยู่ว่า งานใดควรรับบริการอยู่เดี๋ยวนี้ และถ้างานใดที่รับบริการอยู่จะให้รับบริการต่อไปจนเสร็จหรือควรจะเอางานอื่นมาแซงคิวหรือลัดคิวทำงานก่อน

<u>รูปแบบแถวคอย</u> (Queuing Models)

องค์ประกอบพื้นฐานที่เกี่ยวข้องกับรูปแบบแถวคอย คือ ผู้บริการ (servers) แถวคอย (queuing) และ แหล่งที่มา (sources)

คุณสมบัติที่สำคัญของแหล่งที่มา ได้แก่

1) ประเภทของแหล่งที่มา เช่น มีจำนวนจำกัด หรือ มีจำนวนไม่จำกัด

2) ระหว่างเวลาของผู้มาเยือน (inter-arrivaltimes)

3) ความต้องการใช้บริการของงานแต่ละงานในรูปแบบนั้น

คุณสมบัติที่สำคัญของบริการ ได้แก่

1) จำนวนและความจุของแถวคอย

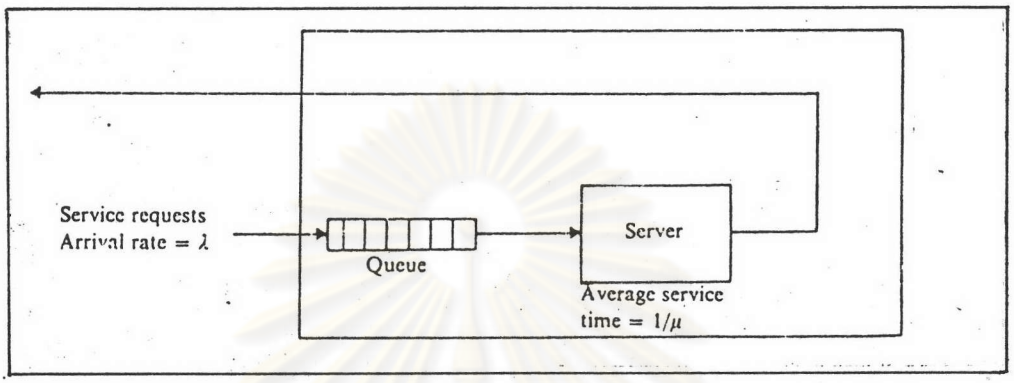
2) จำนวนของผู้บริการ และจำนวนช่องสัญญาณ

3) ความเร็วของผู้บริการ

4) วินัยในการบริการ (Service discipline)

รูปแบบแถวคอยที่ศึกษาประกอบด้วย

1. <u>แถวคอยแบบมีผู้ให้บริการรายเดียว</u> (The single server queue) อธิบายตามรูปที่ 2.5



รูปที่ 2.5 แถวคอยแบบมีผู้ให้บริการรายเดียว

กำหนดให้การร้องขอบริการเข้ามาถึงจุดต่าง ๆ ในเวลา $t_1, t_2 \ldots t_i, \ldots$ และการร้องขอบริการ ซึ่งเข้ามาถึง ณ เวลา t_i ต้องการใช้เวลาในการบริการ s_i หน่วย โดยผู้บริการสามารถให้บริการได้เพียงหนึ่งการร้องขอในขณะใด ๆ กรณีที่ผู้บริการว่างในเวลาที่การร้องขอ ครั้งที่ i เข้ามาถึง การร้องขอจะเข้าใช้บริการทันที และจะออกจากระบบเมื่อใช้เวลา s_i หน่วยแล้ว กรณีที่ผู้บริการไม่ว่างขณะที่การร้องขอเข้ามาถึงการร้องขอนั้นจะถูกบรรจุในแถวคอย เมื่อผู้บริการให้บริการการร้องขอที่มีอยู่เสร็จแล้ว การร้องขอรายการต่อไปจึงถูกเลือกจากแถวคอย และให้บริการจนพอใจ ขั้นตอนวิธีที่ใช้ในการเลือกรายการถัดไปจากแถวคอยเร็วกว่า วินัยในการบริการ ซึ่งวินัยที่ง่ายที่สุดคือ การร้องขอใดเข้ามาก่อนก็รับบริการก่อน (FCFS) หรือการร้องขอใดอยู่ในแถวคอยนานที่สุดก็จะได้รับบริการต่อไป

วินัยอีกประเภท ได้แก่ การร้องขอใดเข้ามาที่หลังจะได้รับบริการก่อน (LCFS) นอกจากนี้การร้องขอในแถวคอยยังต้องการใช้เวลาในการบริการที่แตกต่างกัน วินัยในการบริการอาจเป็น การร้องขอที่ใช้เวลาน้อยสุดจะได้รับบริการ(SRF) หรือการร้องขอที่ใช้เวลามากสุดจะได้รับบริการ (LRF) กรณีที่ไม่ได้กำหนดวินัยในการบริการไว้ชัดเจนจะถือว่าเป็นแบบ FCFS และต้องรู้ว่าแถวคอยมีขนาดเท่าใด ขนาดใหญ่เพียงพอที่จะเก็บการร้องขอทุกรายการไว้ได้หรือไม่

1.1 รูปแบบแถวคอยซึ่งมีแหล่งที่มาไม่จำกัด อัตราซึ่งการร้องขอบริการมาถึงเป็นคุณสมบัติที่สำคัญประการหนึ่งของระบบแถวคอย เพราะง่ายต่อการกำหนดความน่าจะเป็น

$P(t)$ จะต้องมีการร้องขอเข้ามาถึงในช่วงเวลา (t_0, t_0+t) ซึ่งได้ ความสัมพันธ์ คือ

$$P(t) = \lambda * t + O(t)$$

โดยที่ λ คือค่าคงที่ และ $O(t)/t \rightarrow 0$ เมื่อ $t \rightarrow 0$

กำหนดให้การร้องขอที่เข้ามาถึงเป็นการกระจายแบบ ปัวซอง (Poisson)โดยมีความหนาแน่น λ ดังนั้น ความน่าจะเป็นสำหรับการร้องขอรายการจะเข้ามาถึงในช่วงเวลา t คือ

$$P_n(t) = \frac{e^{\lambda * t}(\lambda * t)^n}{n!}$$

ความน่าจะเป็นกรณีที่ไม่มีการร้องขอเข้ามาถึงในช่วงเวลา t คือ

$$P_0(t) = e^{-\lambda * t}$$

ความน่าจะเป็นกรณีที่มีเพียงหนึ่งการร้องขอเข้ามา คือ

$$A(t) = 1 - e^{-\lambda * t}$$

A(t) คือ ความน่าจะเป็นในช่วงเวลาเริ่มต้น t_0 กระทั่งเกิดการร้องขอต่อไปซึ่ง $\leq$ t กรณีที่เริ่มนับเวลาที่มาถึงของการร้องขอที่ i คือ t_i , ดังนั้น A(t) จะแทนค่าความน่าจะเป็นในระหว่างเวลาที่มาถึง$t_{i+1} - t_i$ ซึ่งมีค่าน้อยกว่า t สำหรับการกระจายแบบปัวซอง กำหนดให้การร้องขอระหว่างช่วงเวลาที่มาถึงเป็นตัวแปรแบบสุ่มซึ่งมีอิสระ มีการกระจายแบบเอ็กโปเน็นเชียล (~Exponential distribution) ถ้าคงที่ λ มีอัตราเฉลี่ย เมื่อการร้องขอเข้ามาในระบบ และ $1/\lambda$ คือค่าเฉลี่ยในช่วงเวลาที่เข้ามาถึงดังนั้นบี่ยงเบนมาตรฐานของการกระจายแบบเอ็กโปเน็นเชียลจะเท่ากับค่าเฉลี่ย และเท่ากับ $1/\lambda$

สมมุติฐานสำหรับความน่าจะเป็น P(t) คือ จำนวนการร้องขอเป็นอิสระซึ่งมีอยู่ในแถวคอย ณ เวลานั้น สมมุติฐานจะไม่ถูกต้องกรณีที่เทอร์มินัลมีจำกัด สมมุติฐานของความน่าจะเป็นของเวลาให้บริการที่มีการกระจายแบบเอ็กโปเนียนเชียลกำหนดให้ s_i คือตัวแปรที่มีการกระจายอิสระแบบสุ่ม ดังนั้น การกระจายแบบเอ็กโปเน็นเชียล คือ

$$B(t) = 1 - e^{-\mu * t}$$

B(t) คือ ความน่าจะเป็น ซึ่งการร้องขอบริการเลือกเข้ามาแบบสุ่ม และต้องการเวลาไม่เกิน t หน่วย สำหรับการร้องขอ เวลาเฉลี่ยที่ให้บริการต่อหนึ่งการร้องขอคือ $1/\mu$ ซึ่งเท่ากับค่าเบี่ยงเบนมาตรฐานของการกระจายของเวลาให้บริการ

สำหรับแถวคอยแบบมีผู้บริการรายเดียว การกระจายในระหว่างช่วงเวลาที่เข้ามาถึง และการกระจายของเวลาให้บริการ กำหนดอัตราการให้บริการได้ดังนี้

$$\rho = \frac{7^{\lambda}}{\mu}$$

กำหนดให้การร้องขอบริการเข้ามาถึงด้วยอัตราเฉลี่ย λ การร้องของตัวหน่วยเวลา และต้องการเวลาบริการเฉลี่ย $1/\mu$ หน่วยเวลา ดังนั้น แถวคอยอาจมีขนาดใหญ่ ถ้า λ/μ มีค่ามากกว่า 1 จึงอาจสรุปได้ว่า ρ น้อยกว่า 1 สำหรับแถวคอยแบบมีผู้ให้บริการรายเดียวที่มีการกระจายของการร้องขอที่มาถึงแบบปัวซอง และการกระจายของเวลาการให้บริการแบบเอ็กโปเน็นเชียล สามารถบอกสถานะทฤษฎีแถวคอย ได้ว่าระบบจะถึงสถานะพร้อม ตามจำนวนเฉลี่ยของการร้องขอในระบบกำหนดเป็นสูตร คือ

$$\bar{n} = \frac{\rho}{1-\rho} \qquad \text{(ก)}$$

เมื่อเวลาเฉลี่ยที่การร้องขอต้องการ คือ $1/\mu$ ดังนั้น ค่าเฉลี่ยของเวลาตอบสนองในแถวคอยแบบหนึ่งบริการ คือ

$$W = \frac{1}{\mu} + \frac{1}{\mu} * \frac{\rho}{(1-\rho)} \qquad \text{(ข)}$$

แถวคอยแบบ M/M/1/∞ โดยสัญญลักษณ์ตัวที่หนึ่งหมายถึงการประมวลผลอินพุท ตัวที่สองคือการกระจายของบริการ ตัวที่สามคือจำนวนผู้ให้บริการ และสัญญลักษณ์ตัวที่สี่คือ รูปแบบแหล่งที่มาซึ่งไม่จำกัด ทั้งนี้วินัยในการให้บริการ หากไม่ได้กำหนด และถือว่าเป็นแบบ FCFS ดังนั้น M ตัวที่หนึ่งหมายถึง การประมวลผลอินพุท แบบปัวซอง M ตัวที่สองหมายถึง การให้บริการซึ่งมีการกระจายแบบเอ็กโปเน็นเชียล และ 1 คือ แถวคอยแบบมีผู้ให้บริการรายเดียว ให้ E คือ การกระจายแบบเออแลงก์ (Erlang distribution)

แถวคอยแบบ M/G/1 คือ แถวคอยที่มีผู้ให้บริการรายเดียว มีการประมวลผลการเข้ามาแบบปัวซอง มีเวลาให้บริการแบบความน่าจะเป็น โดยมีค่าเฉลี่ยคือ $E(S) = 1/\mu$ และค่าเบี่ยงเบนมาตรฐาน คือ $\delta(S)$ ของการกระจายเวลาให้บริการที่รู้ได้ และสมมุติว่า $\rho = \lambda/\mu < 1$ และกำหนด c ได้ดังนี้

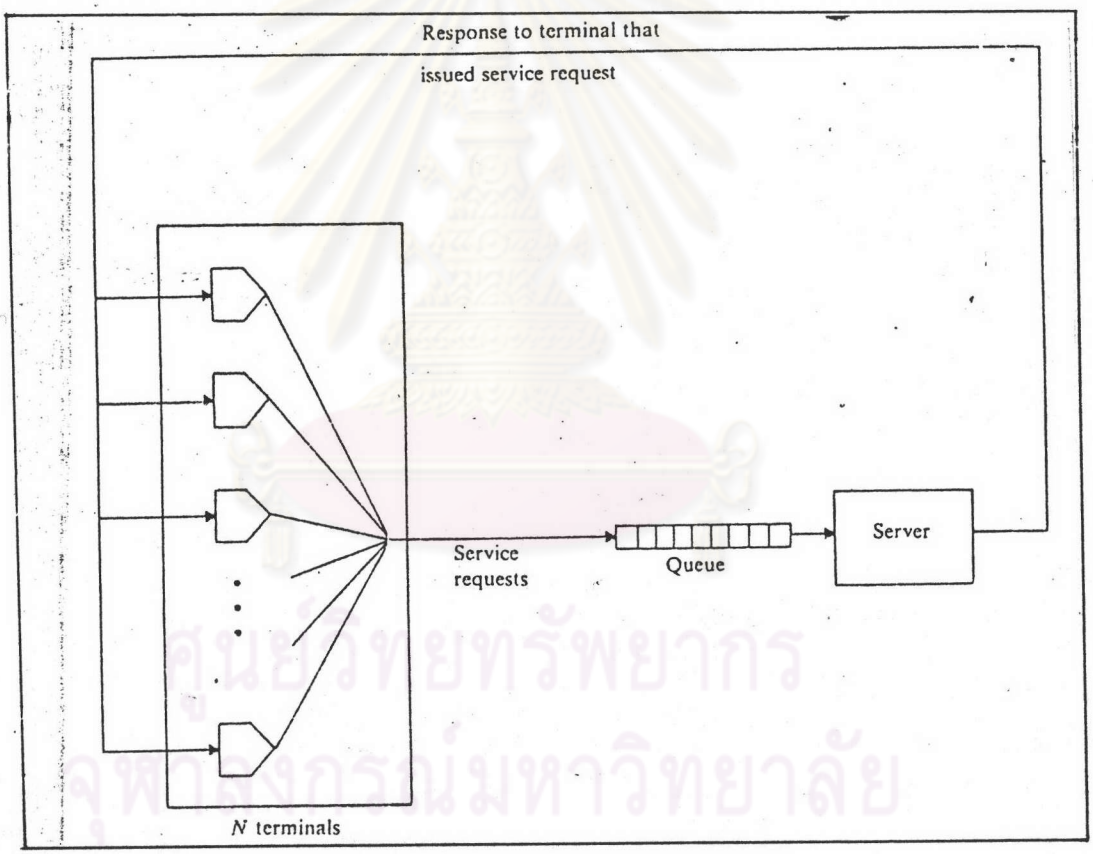
$$c = \frac{\delta^2(s)}{E^2(s)}$$

ซึ่งเขียนเป็นระบบแถวคอยที่ถึงสถานะพร้อมตามจำนวนเฉลี่ยการร้องขอในระบบเป็นสูตรดังนี้

$$W = \frac{1}{\mu} + \frac{1}{\mu} * \frac{\rho}{(1-\rho)} * \frac{(1+c)}{2} \qquad \text{(ค)}$$

สำหรับการกระจายเวลาการให้บริการแบบเอ็กโปเน็นเชียล $\delta(S) = E(S)\ c$ จะเท่ากับ 1 และค่าเฉลี่ยเวลาตอบสนองตามสูตร (ข) จะเท่ากับ (ก)

1.2 รูปแบบแถวคอยซึ่งมีแหล่งที่มาจำกัด ตามรูปที่2.6 โดยทั่วไประบบจะมีจำนวนเทอร์มินัลแน่นอนตามจำนวนผู้ใช้ กำหนดให้ เท่ากับ N ในบางระบบอาจไม่สมเหตุสมผลที่จะกำหนดว่าอัตราการมาถึงของการร้องขอบริการมีจำนวนการร้องขออิสระซึ่งพร้อมอยู่ในแถวคอย ในกรณีที่เกินความคาดคิด คือ มีการร้องขอ บริการเข้ามาในระบบ จากจำนวนเทอร์มินัลทั้งหมด N หน่วยจะไม่มีการร้องขอใหม่จนกว่ามีอย่างน้อยหนึ่งการร้องขอได้รับบริการเสร็จ ระบบประเภทนี้จะสะสมการร้องขอไว้ในแถวคอยทำให้การมาถึงของการร้องขอใหม่ ลดลง



รูปที่ 2.6 แถวคอยซึ่งมีแหล่งที่มาจำกัด

สมมุติฐานของแถวคอยที่มีแหล่งที่มาจำกัด คือ แต่ละเทอร์มินัลในสถานะผู้ใช้ ซึ่งมีความน่าจะเป็น P(t) เทอร์มินัลจะส่งการร้องขอในช่วงเวลา (t_0, t_0+t) กำหนดความสัมพันธ์ได้ดังนี้

$$P(t) = \lambda * t + O(t) \text{ โดย } t \text{---}>0 \text{ และ } \lambda \text{ คือค่าคงที่}$$

จากค่า λ และ P(t) เป็นอิสระ ณ t_0 จากสมมุติฐานของมาคอร์ฟ สำหรับแบบแหล่งที่มาจำกัด และการกระจายของเวลาบริการ เป็นแบบเอ็กโปเน็นเชียลรูปแบบแถวคอยคือ M/M/1/N ซึ่งเขียนเป็นสูตรค่าเฉลี่ยสถานะพร้อม มีเวลาตอบสนอง คือ W

$$W = \frac{N/\mu - 1}{1-p_0 \quad \lambda}$$

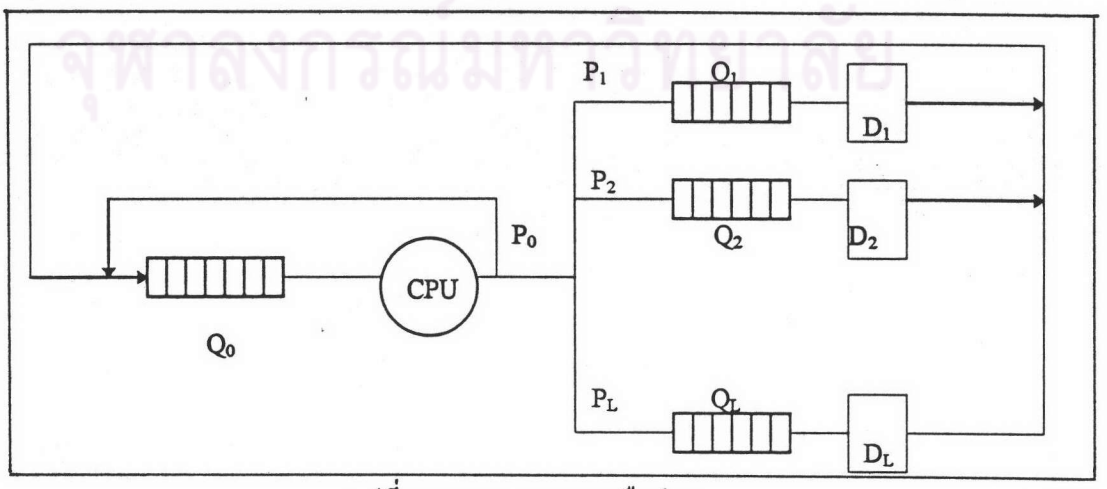
เมื่อ p_0 คือเศษส่วนของเวลาทั้งหมด โดยที่การประมวลผลแบบแบ่งเวลาว่างปริมาณ $(1- p_0)$ คือ ความน่าจะเป็น โดยที่กระบวนการทำงาน p_0 หาได้จากสูตร

$$\sum_{i=0}^{N} Pi = 1 \quad \text{และ} \quad P_n = \frac{N!}{(N-n)!} * (\frac{\lambda}{\mu})^n * P_0$$

โดยที่ $1/\mu$ คือเวลาบริการ E(S) และ $1/\lambda$ คือค่าเฉลี่ยเวลาของผู้ใช้ หรือเวลาที่คิด (Think Time) E(U) ดังนั้น สูตรเวลาตอบสนองคือ

$$W = \frac{N * E(S) - E(U)}{1 - p_0}$$

2 <u>แถวคอยของเครือข่าย</u> (Networks of queues) ตามรูปที่ 2.7 การศึกษากำหนดให้มี N โปรแกรมในระบบ ซึ่งมี 1 ซีพียู และ อุปกรณ์รอบข้าง L หน่วย และโปรแกรมเริ่มจากจบแถวคอยของซีพียู เมื่อโปรแกรมอยู่ที่หัวแถวคอยก็จะได้บริการโดยมีการกระจายเวลาบริการด้วยค่าเฉลี่ย $1/\mu_0$ นั้นคือมีความน่าจะเป็น p_0 โปรแกรมจะสิ้นสุด โดยสมมุติต่อไปว่าจะมีโปรแกรมใหม่เข้ามาทันที



รูปที่ 2.7 แถวคอยของเครือข่าย

ที่สิ้นสุดแถวคอยซีพียู ดังนั้นโปรแกรมจำนวน N หน่วยจะอยู่คงที่ในระบบ ในกรณีที่โปรแกรมไม่สิ้นสุด ความน่าจะเป็น คือ $p_1, p_2, \ldots, p_L$ เข้าไปในแถวคอย $Q_1, Q_2, \sim\ldots, Q_L$ ตามลำดับ เมื่อโปรแกรมอยู่ที่หัวของ Q_i ก็จะได้รับบริการจากอุปกรณ์รอบข้าง Di โดยการกระจายเวลาบริการด้วยค่าเฉลี่ย $1/\mu_i$ จากนั้น โปรแกรมจึงเข้าไปประมวลผลในหน่วยประมวลผลกลางจากความน่าจะเป็น $p_1, p_2, \ldots, p_L$ มักจะกำหนดว่า

$$p_1 + p_1, +\ldots. +p_L = 1$$

ในแถวคอยเครือข่ายการกระจายของเวลาให้บริการ เป็นแบบเอ็กโปเน็นเชียลโดยทั่วไป กำหนดให้มีผู้ใช้ N ราย ในเครือข่าย และมีการติดต่อระหว่างผู้ให้บริการซึ่งเชื่อมโยงกัน (sources) L+1 หน่วย ผู้ให้บริการแต่ละรายมีบริการแถวคอยที่สัมพันธ์กัน กำหนดให้ p_{ij} คือ ความน่าจะเป็นที่ผู้ซื้ออยู่ในผู้ให้บริการที่ i และเข้าไปใช้บริการแถวคอยที่ผู้ให้บริการที่ j ดังนี้ p_{ij} ทุกตัวมีค่า ≤ 1 สำหรับ i ทุกตัว ดังนั้น

$$\sum_{j=0}^{L} p_{ij} = 1$$

รูปแบบแถวคอยเครือข่ายกำหนดภายใต้สมมุติฐาน คือ ผู้ให้บริการทุกรายมีการกระจายของเวลาการให้บริการแบบเอ็กโปเน็นเชียลอิสระ ความน่าจะเป็นที่สถานะพร้อม $p(n_0, n_1, \ldots n_2)$ สำหรับผู้ใช้จำนวน n_j ราย ณ ผู้ให้บริการตัวที่ j ดังนั้น

$$0 \leq n_j \leq N \text{ และ } \sum_{j=0}^{L} n_j = N$$

ซึ่งสามารถเขียนเป็นสมการง่าย ๆ โดยการ $y_0, y_1, y_2, \ldots y_L$ เป็นการเลือกใด ๆ

$$y_j = \sum_{i=p}^{L} p_{ij} y_i \qquad \text{สำหรับ } j = 0,1,\ldots,L$$

ฉะนั้นความน่าจะเป็นที่สถานะพร้อม คือ

$$P(n_0, n_1, \ldots n_L) = \frac{1}{G} \prod_{k=0}^{L} -(\frac{y_k}{\mu_k})^{nk}$$

โดยที่ $1/u_k$ คือค่าเฉลี่ยเวลาบริการสำหรับผู้ให้บริการที่ k และ G คือ ค่ามาตรฐานที่ทำให้ผลรวมของค่า $p(n_0, n_1, n_2.----n_L)$เท่ากับ 1 ดังนั้น

$$G = \sum_{k=0}^{L} \pi(\frac{y_k}{\mu_k})^{nk}$$

โดยผลรวมเกิดจากค่าทั้งหมดของกลุ่ม $((n_0, n_1, n_2.----n_L)$ ที่พอใจ ซึ่งเป็นสูตรง่าย ๆ สำหรับผู้ให้บริการแบบส่วนกลาง

<u>วิธีการเลือกและกำหนดตัวแปร</u>

ในการเลือกและกำหนดตัวแปรสำหรับการประเมินความสามารถของระบบคอมพิวเตอร์ สามารถกำหนดได้จากชนิดของภาระงาน อรรถประโยชน์ และระดับการบริการการพิจารณาถึงลักษณะการทำงานของระบบ สามารถแบ่งออกได้เป็น 3 ส่วน ดังนี้

1. <u>ภาระงาน</u> (Work Load) โดยมีหน่วยในการวัดและตรวจสอบเป็นจำนวนงานต่อชั่วโมง สำหรับงานแบบแบทช์ หรือจำนวนทรานแซกชั่นต่อวินาที สำหรับงานแบบการดึงข้อมูลระบบแฟ้มข้อมูลหรือฐานข้อมูล หรือจำนวนอินเตอร์แอกชั่นต่อวินาที สำหรับงานแบบการติดต่อสอบถามกับเครื่องโดยตรง

2. <u>อรรถประโยชน์ของทรัพยากรในส่วนของฮาร์ดแวร์</u> อันได้แก่ อรรถประโยชน์ซีพียูช่องสถานี จานแม่เหล็ก เทป และเครื่องพิมพ์

3. <u>ระดับการบริการแก่ผู้ใช้บริการ</u> ถ้าเป็นแบบออนไลน์เป็นจำนวนเวลาที่ตอบสนองคำถามกลับมา หรือถ้าเป็นแบบ เป็นจำนวนเวลาตั้งแต่สั่งงานจนได้รับงานจากระบบคอมพิวเตอร์

การกำหนดตัวแปรตามโปรแกรมควบคุมระบบ การพิจารณาโปรแกรมควบคุมระบบตามหน้าที่ของการทำงาน (รายละเอียดอธิบายในบทที่ 3)

การกำหนดตัวแปรตามข้อมูลตามความต้องการของบุคลากรในหน่วยงานคอมพิวเตอร์ ซึ่งสามารถแบ่งได้เป็น 5 กลุ่ม คือ

1) ฝ่ายปฏิบัติการคอมพิวเตอร์
2) ฝ่ายโปรแกรมเมอร์ระบบคอมพิวเตอร์
3) ฝ่ายผู้ใช้บริการ
4) ฝ่ายพัฒนาโปรแกรมงานต่าง ๆ
5) ผู้บริการหน่วยงาน

ตัวอย่างเช่น ข้อมูลที่จำเป็นสำหรับฝ่ายปฏิบัติการคอมพิวเตอร์จะมีข้อมูลเป็นดังนี้

1) จำนวนเวลาของหน่วยประมวลผลการที่สามารถใช้ได้ต่อวัน
2) จำนวนเวลาของหน่วยประมวลผลการที่ถูกใช้ไปจริงต่อวัน
3) จำนวนเวลาของหน่วยประมวลผลการที่ถูกใช้งาน
4) อรรถประโยชน์ของทรัพยากรเป็นช่วงของเวลาต่อวัน
5) จำนวนเวลาทั้งหมดที่งานแบบแบทช์ที่สำคัญใช้
6) เวลาการตอบสนองการถามตอบข้อมูลที่เทอร์มินัลและหน่วยประมวลผลกลาง
7) ปริมาณทรัพยากรที่จำเป็นต้องใช้สำหรับงานที่สำคัญ
8) ปริมาณทรัพยากรที่จำเป็นต้องใช้โดยเฉลี่ยต่องาน