Control de congestión en los extremos en comunicaciones de granularidad fina

Arquitectura y Computación de Altas Prestaciones

Vladislav Nikolov Vasilev

Índice

Comentar los contenidos que se van a tratar.

1. Introducción

Redes HPC se caracterizan por la gran cantidad de tráfico y porque son redes sin pérdidas.

Tipos de tráfico:

- **Admisible**: tráfico dirigido a cada extremo de la red que no requiere más recursos de los disponibles.
- **No admisible**: tráfico dirigido a cada extremo de la red que requiere más recursos de los disponibles.

Ambos tipos de tráfico pueden causar congestión de árbol.

Uso de GPUs en HPC en alza → mensajes pequeños (comunicaciones de granularidad fina).

Las redes en HPC son redes, tal y como su propio nombre indica, de alto rendimiento. Se caracterizan por tener una gran cantidad de tráfico y porque son redes sin pérdidas, esto es, no se pueden perder los mensajes que se envían. Por tanto, la congestión en este tipo de redes es un gran problema, ya que el rendimiento se reduce.

Los autores clasifican el tráfico en dos categorías. COMENTARLAS. Cabe destacar que el tráfico admisible puede llegar a generar congestión de fábrica, que es un tipo de congestión generada por mal enrutamiento en la que se satura un nodo porque este recibe más tráfico.

Ambos tipos de tráficos pueden causar congestión de árbol. La congestión de árbol consiste en lo siguiente: cuando un nodo tiene su buffer de entrada lleno le indica a los nodos que le envían información que dejen de hacerlo, ya que este no tiene capacidad para guardar información nueva. Por tanto, la información se comienza a acumular en los buffers de entrada de dichos nodos, congestionándolos también y obligándolos a dejar de recibir nueva información. De esta forma se puede llegar a congestionar y relantizar toda la red de forma muy rápida si no existe ningún mecanismo de control.

En la actualidad en HPC se están usando cada vez más las GPUs. Para comunicarse y acceder a recursos compartidos, las hebras envían mensajes, los cuáles son pequeños.

2. Mecanismos existentes de control de congestión

- Software: algoritmos de enrutamiento adaptativos.
- Hardware:
  - ECN (*Explicit Congestion Notification)*. Usado en conexiones Infiniband. Control reactivo.
  - SRP (*Speculative Reservation Protocol*). Control proactivo.

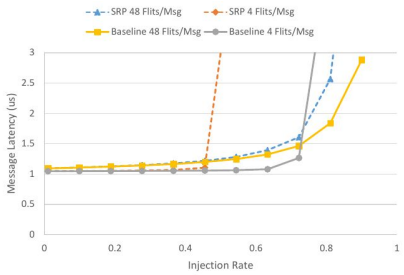


Figure 2: Comparison of SRP's performance on medium and small messages.

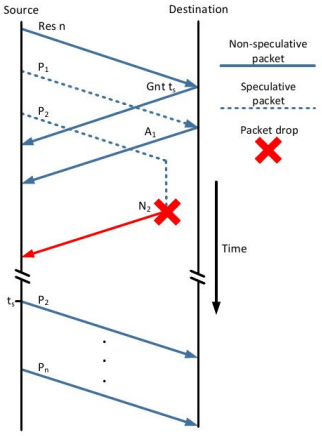


Figure 1: SRP operation diagram

Existen algunos mecanismos para el control de congestión.

- Por software: tenemos los algoritmos de enrutamiento adaptativos, los cuáles permiten reducir la congestión de fábrica al enrutar mejor los datos.
- Por hardware:
  - Por una parte está ECN, el cual se utiliza en conexiones Infiniband. Cuando se produce una congestión, se manda una notificación a toda la red, reduciendo el tráfico de la red. Es un mecanismo reactivo, ya que espera a que se produzca la conexión.
  - Por otra parte está el protocolo SRP, el cual hace un control proactivo de la congestión, evitando que se forme. (SE EXPLICA EL PROTOCOLO CON LA FIGURA).

En la figura se observa que el protocolo funciona muy bien para mensajes grandes pero bastante mal para los pequeños. El problema está en los mensajes de reserva, ya que estos acaparan la mayor parte de los recursos de la red. Por tanto, los mensajes como tal, a pesar de ser pequeños, tienen menos recursos disponibles y deben esperar más tiempo para poder llegar a su destino.

3. Nuevas propuestas

SMSRP (*Small-Message Speculative Reservation Protocol*)

- Basado en el protocolo SRP.
- No es necesario hacer la reserva siempre → reservar solo en caso de congestión.
- Fácil de implementar en hardware si SRP está implementado.
- Si se empiezan a descartar muchos paquetes la red se va a comenzar a llenar de mensajes de reserva.

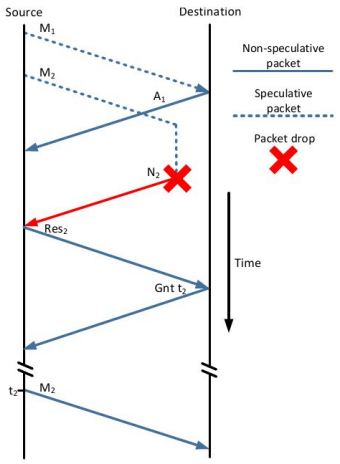


Figure 3: SMSRP operation diagram.

La primera propuesta es SMSRP, la cuál se basa en SRP. Los autores se dieron cuenta que no hace falta realizar una reserva siempre, si no que basta hacerla cuando se detecta una congestión.

(EXPLICAR PROTOCOLO CON FIGURA)

La principal ventaja es que es muy sencillo de implementar en hardware si se tiene SRP implementado, ya que basta con reordenar el orden en el que se envían los paquetes especulativos y la reserva. El problema principal es que si se empiezan a descartar muchos mensajes, la red se va a comenzar a llenar de mensajes de reserva, los cuáles van a saturar la red. Por tanto, tendría un comportamiento parecido a SRP.

3. Nuevas propuestas

LHRP (*Last-Hop Reservation Protocol*)

- El switch final antes del extremo se encarga de las reservas.
- Se eliminan mensajes de reserva enviados por el origen.
- Solo el último switch puede descartar paquetes.
- Último switch tiene un umbral que indica el número máximo de paquetes en la cola de entrada. Cuando se supera el umbral, se comienzan a descartar los paquetes.

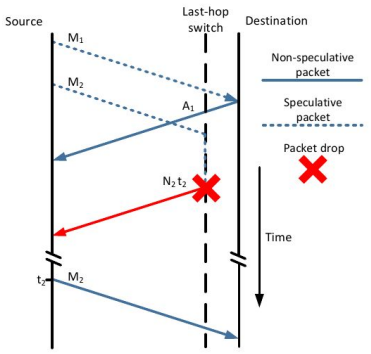


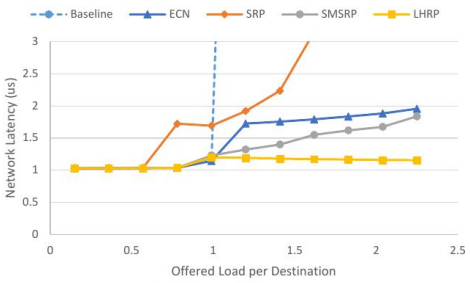
Figure 4: LHRP operation diagram.

La segunda propuesta es el protocolo LHRP.

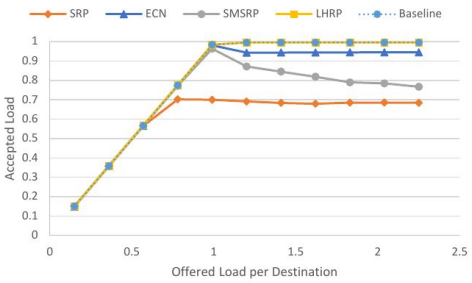
(EXPLICAR PROTOCOLO CON LA FIGURA Y LOS PUNTOS)

4. Experimentación y resultados

Pruebas con red donde hay tráfico constante dirigido hacia unos pocos extremos, simulando la congestión en estos.



(a) Network latency



(b) Data throughput

7

Comentar el entorno utilizado: simulador donde se tiene una red Dragonfly de 1056 nodos con mensajes de 4 flits. En este caso, 60 nodos envían mensajes a 4 extremos, de forma que se simula la congestión en los extremos.

Figura de la izquierda: LHRP apenas genera latencia de red pasado el punto de saturación, de forma que casi no se produce congestión de ningún tipo en la red. SMSRP tiene una latencia de red algo mayor que LHRP a medida que va aumentando el tráfico pasado el punto de saturación debido a que se empiezan a descartar más mensajes especulativos y se envían más mensajes de reserva, lo cuál sobrecarga ligeramente la red. A pesar de eso, son los protocolos que mejores resultados han obtenido si los comparamos con ECN, SRP y una red sin control de congestión.

Figura de la derecha: Pasado el punto de saturación, LHRP no genera sobrecarga en la red, de forma que casi la totalidad de los mensajes que viajan por la red son de datos, al igual que ocurre en una red sin control de congestión. ECN está cerca de este nivel (notificaciones viajan por la red). En el caso de SRP, a partir de dicho punto el 30 % del tráfico es de mensajes de reserva. A medida que va aumentando el tráfico, el comportamiento del protocolo SMSRP empieza a parecerse al de SRP, ya que cada vez habrá más mensajes de reserva viajando por la red.

4. Experimentación y resultados

Pruebas con red donde aparece mucho tráfico repentinamente.

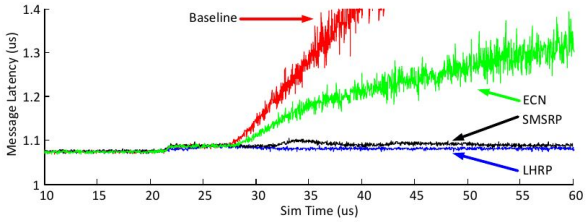


Figure 6: Network transient response to the onset of congestion.

Con SMSRP y LHRP la latencia de los mensajes es pequeña, siendo la latencia del LHRP ligeramente menor que la de SMSRP. Por tanto, en estos casos se puede afirmar que no se produce ningún tipo de congestión de la red. Los otros dos mecanismos probados fueron ECN y la red sin control de congestión. En ellos se obtuvieron unos peores resultados, pudiéndose apreciar la aparición de congestión.

4. Experimentación y resultados

Pruebas con red sin congestión. Estudio de la sobrecarga que produce el control de congestión.

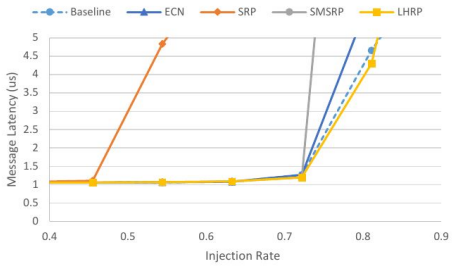


Figure 7: Network performance of uniform random traffic with 4-flit messages.

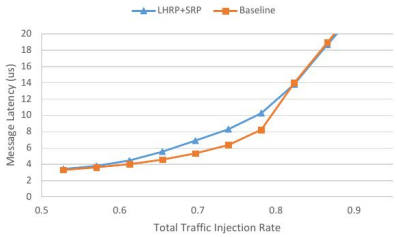
Tomando como referencia una red sin control de congestión, se vio que a medida que aumenta el tráfico las dos propuestas se quedaban bastante cerca de la red sin control de congestión, indicando por tanto que en líneas generales producen poca sobrecarga. ECN también se quedaba relativamente cerca, mientras que SRP se quedaba más atrás, lo cuál implica que es el protocolo que genera una mayor sobrecarga en la red.

5. Modificaciones sobre las propuestas

Modificaciones sobre LHRP:

- Descartar mensajes en otros puntos del camino que no sean el switch final.
- Combinar LHRP con SRP y elegir uno de los dos protocolos en función del tamaño del mensaje si se quieren enviar los dos tipos de mensajes en la misma red.
- Combinar enrutamiento adaptativo con LHRP.

(a) 4-flit message latency



(b) 512-flit message latency

Respecto al primer punto: De esta forma, el rendimiento mejora si se da el caso en el que hay mucha congestión en el extremo.
Respecto al segundo punto: Se puede también modificar el umbral de la cola en el último switch en función del tamaño de los mensajes que se van a enviar.

6. Conclusiones

El control de congestión en los extremos requiere de mecanismos proactivos, rápidos y que generen muy poca sobrecarga.

LHRP y SMSRP han mostrado un buen comportamiento para mensajes pequeños.

Se puede combinar LHRP con SRP en redes donde se envían mensajes pequeños y grandes.

Comentar las conclusiones que se pueden extraer.