Control de congestión en los extremos en comunicaciones de granularidad fina

Arquitectura y Computación de Altas Prestaciones

Vladislav Nikolov Vasilev

Índice

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).

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.

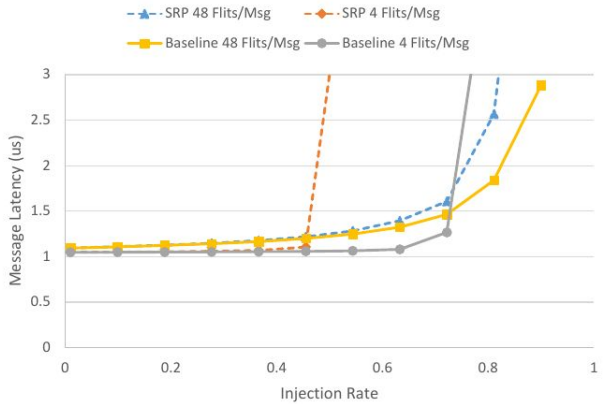


Figure 2: Comparison of SRP's performance on medium and small messages.

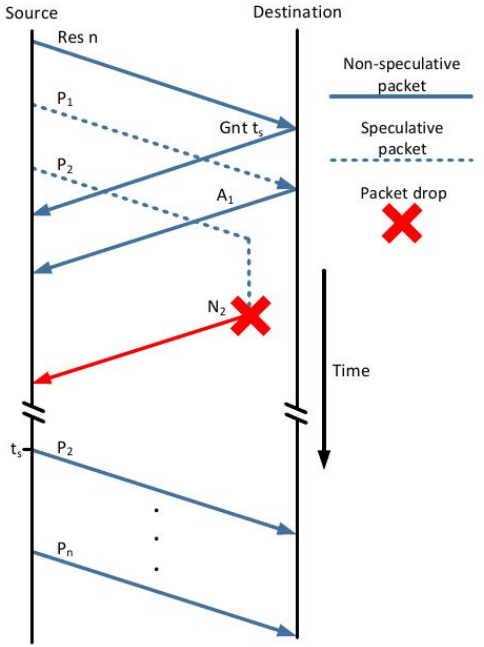


Figure 1: SRP operation diagram

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.

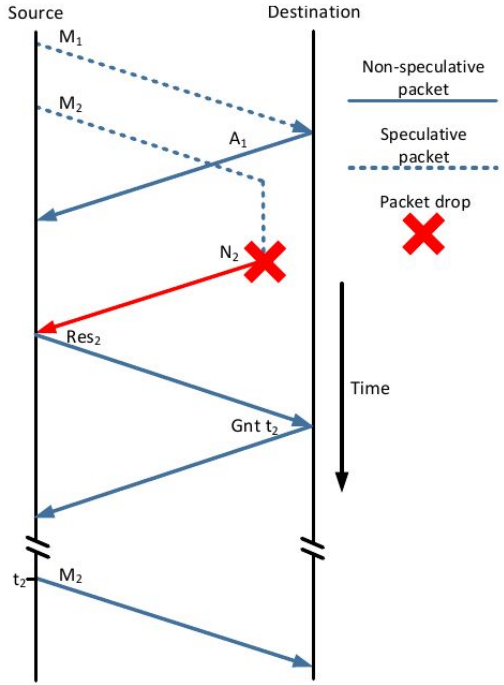


Figure 3: SMSRP operation diagram.

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.

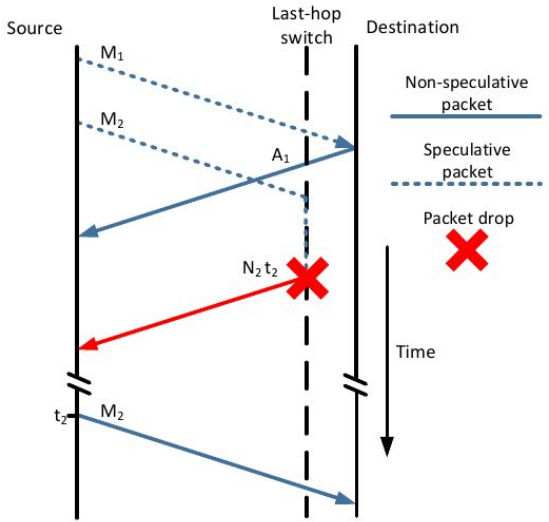
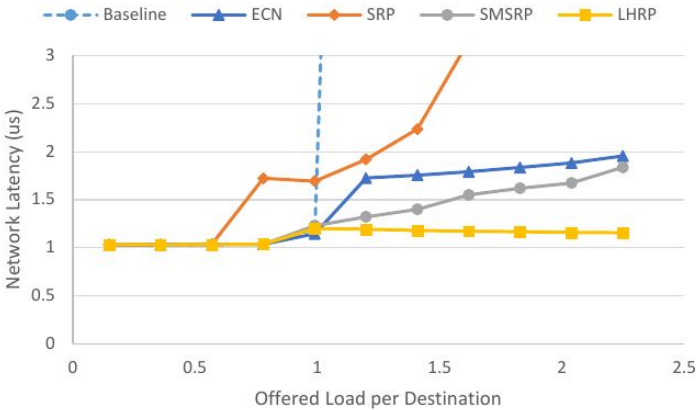


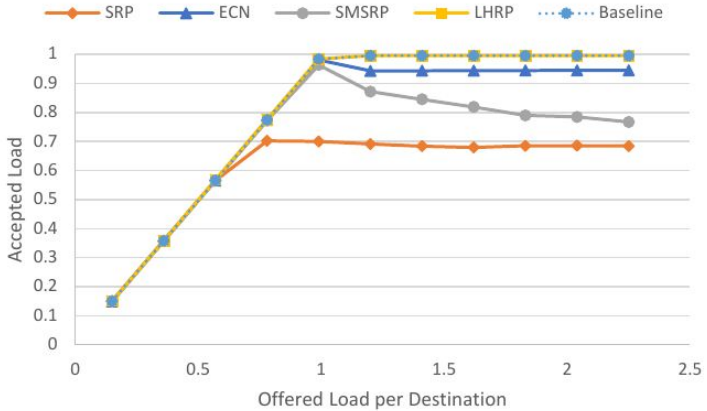
Figure 4: LHRP operation diagram.

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

4. Experimentación y resultados

Pruebas con red donde aparece mucho tráfico repentinamente.

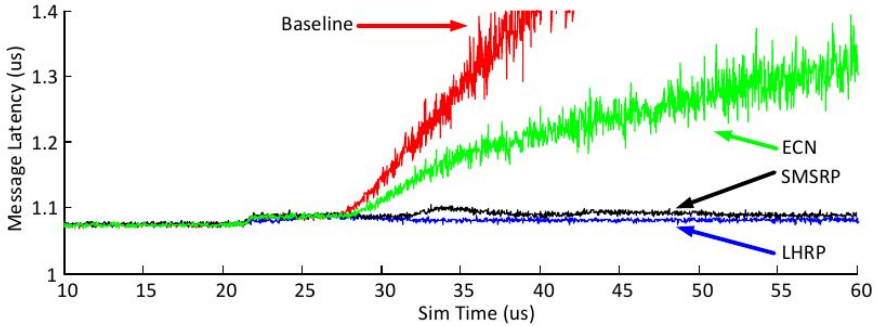


Figure 6: Network transient response to the onset of congestion.

4. Experimentación y resultados

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

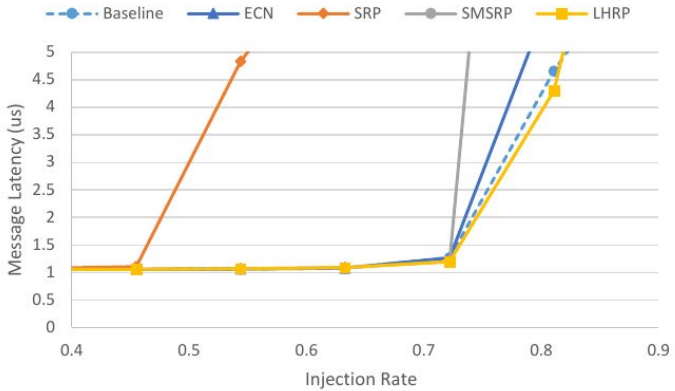
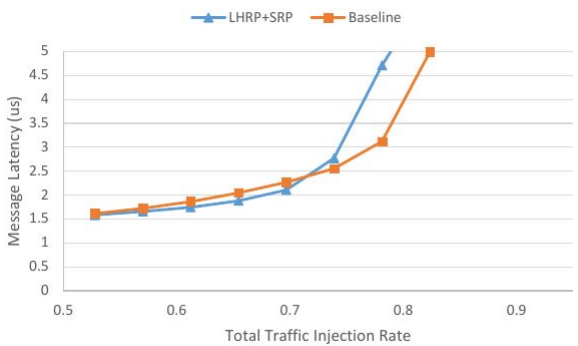


Figure 7: Network performance of uniform random traffic with 4-flit messages.

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

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.